**ELECTRONIC, ELECTRICAL & SYSTEMS ENGINEERING**

# UNIVERSITY OF BIRMINGHAM

BEng Mechatronic and Robotics Engineering

Final Year Project (EE3P)

Bauan Rashid

1946857

Supervisor: Dr Neil Cooke

## A Comparative Analysis of state-of-the-art Pre-Trained Face Recognition Models and the new Face Transformer on the LFW Dataset

# Project Self Assessment

**Checklist: Put a "Y" in the column which corresponds to your assessment of our own ability**

| Category | I find this very difficult | I find this a bit difficult | neutral | I find this fairly easy | I find this very easy |
|---|---|---|---|---|---|
| Ability to work independently | | | | | Y |
| Ability to manage my time | | | | | Y |
| Ability to learn new skills or concepts in depth | | | Y | | |
| Ability to learn new concepts or skills quickly | | | Y | | |
| Ability to focus on targets | | | | Y | |
| Ability to apply things that I have learned | | | | Y | |
| Ability to understand the implications of results and findings | | | | Y | |
| Ability to draw conclusions | | | | | Y |

Comment on your self assessment checklist. To obtain full marks you must explain the reasons why you selected particular columns for each of your ability criteria.

| | |
|---|---|
| What aspects of your project did you enjoy and or went well (up to 50 words) | Since I am analysing different models, it is important to dig deep into understanding what the strengths and weaknesses of the models are and it became clear that the face transformer, though still young in development, had major flaws in being able to perform face recognition under different stress tests. |
| What aspects of your project did you find difficult or would you change (up to 50 words) | The pre-trained models were trained and used for face classification while I was testing it on face verification which is different. Changing and standardising all the tests and models proved challenging especially for the face transformer. I would focus more on trying to understand the code of the models before implementing them. |

# List of Contents

## List of Figures

### List of Tables



# Tangible Deliverables

All documents, code and a summary of the Tangible Deliverables of this study will be both uploaded on Canvas and available through the GitHub repository link [Cleary. 2023] found in the references. All code for analyses, plots , and Image alterations have been written by myself excluding the code of the pre-trained models found in [Timesler. (2019)] and [Zhong. 2021].

# Acknowledgements

I want to thank everyone in my family for their support throughout my University. Semester 2 has been a lot of work, and much of my part of the house chores were done by my brothers and sisters and of course a special thank you to my mother who exerts all her energy for our sake.
I also want to thank my supervisor for this project, Dr Neil Cooke, for the advice in pushing the project into the right direction when things were unclear.

# Abstract

Face recognition technology has seen significant advancements in recent years. There are 3 objectives to this study: to evaluate and compare the InceptionResNetV1 with the newly released face transformer, test how different training datasets affect the performance of the InceptionResNetV1, and provide an in-depth analysis of the face transformer. For the first objective, the InceptionResNetV1 performed much better in comparison to the Face transformer in most tests. For the second objective, the InceptionResNetV1 trained on the VGGFace2 dataset displayed better performance and robustness compared to the InceptionResNetV1 trained on the Casia-Webface dataset. This is due to VGGFace2 having 8 times more images than the Casia-Webface, increasing the diversity of training data. Lastly, for the third objective, the face transformer showed potential in addressing occlusion challenges. However, it had poor generalising capabilities for low resolution images and filters. The implications of this study showcase the importance of model fine-tuning and dataset selection for optimal performance of face recognition models under diverse conditions. This study also makes clear the importance of exercising repeatability, controllability, and transparency which was lacking in the face transformer paper.

# 1: Introduction

## 1.1: Background and Motivation

Face Recognition technology has integrated into many aspects of daily life which has increased the demand for reliable and efficient systems used in important applications such as security and surveillance. Despite significant advancements in this field, challenges remain in achieving accurate, fast and robust face recognition under varying conditions. This research aims to address these challenges by investigating the performance of different face recognition models under diverse conditions that will be tested on the LFW dataset.

This research aims to contribute to the development of reliable and accurate face recognition systems and for researchers to analyse how the face transformer performs in comparison to proven state-of-the-art models.

## 1.2: Research Objectives

The primary objectives of this research are shown in Figure 1:

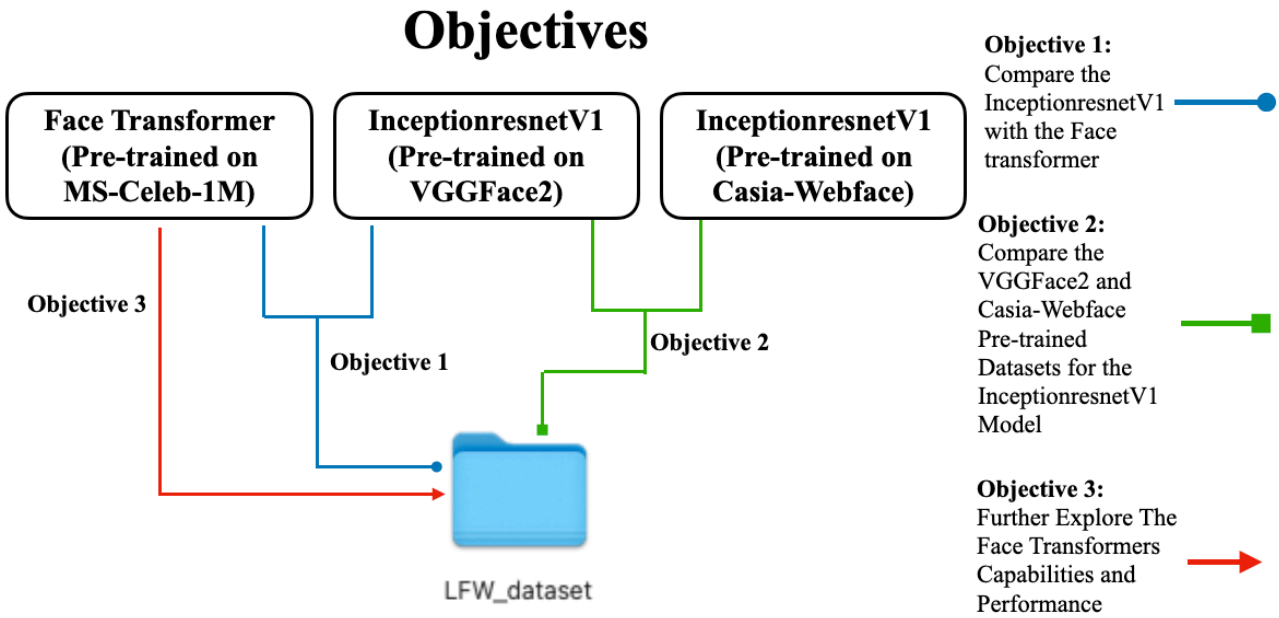


Figure 1: Overview of the objectives in this study

- Objective 1: Compare the InceptionResNetV1(VGGFace2) model against the face transformer model for face recognition under occlusion, image alterations, varying lighting and filters.
- Objective 2: Evaluate the change in performance when the InceptionResNetV1 is trained on 2 different datasets namely the VGGFace2 and Casia-Webface - both popular training datasets for face recognition.
- Objective 3: In-depth analyses for the newly released face transformer with its unique patch-based approach and self-attention mechanisms.

## 1.3: Hypothesis

This research will test the following hypotheses:

- Hypotheses 1: Pre-trained models like MTCNN and InceptionResNetV1 offer superior face recognition performance compared to other models like the new face transformer.
- Hypotheses 2: The VGGFace2 Dataset with 3.31 million images of 9,131 subjects will display better robustness to stress tests than the Casia-Webface with 494,414 images of 10,575 subjects due to a more diverse set of images, even when using identical architectures and undergoing the same evaluation tests.
- Hypotheses 3: The face transformer (employing a patch-based approach to concentrate on local facial features and their relationships) will demonstrate improved robustness in the occlusion tests in comparison to traditional models that predominantly rely on global facial features.

## 1.4: Scientific Principles

For experimental analysis, the paper "Sustainable computational science: the ReScience initiative" by [Rougier, N. P et.al 2017], gives a comprehensive discussion on the principles of reproducibility. The key takeaways that have been used in this study are:
Repeatability: when another researcher can repeat all experiments without any issues.
Controllability: keeping all variables the same except the ones being changed for the experiment to assure fair analysis.
Transparency: all the methods, data, and tools used in the study should be clearly documented to ensure other researchers can understand what was done, why it was done, and how to reproduce it.
These scientific principles will be reiterated throughout the study to showcase its adherence and importance.

## 1.5: Project Management

For day-to-day progress and tracking research, the notion app was used to write detailed notes that was then transferred into the log book at the end of the day. Including note taking, the notion app was very convenient as it gives historic changes of the work. It is easily editable and collectable for the report and provided a gantt chart update with each changes made for this project as shown in Figure 2.

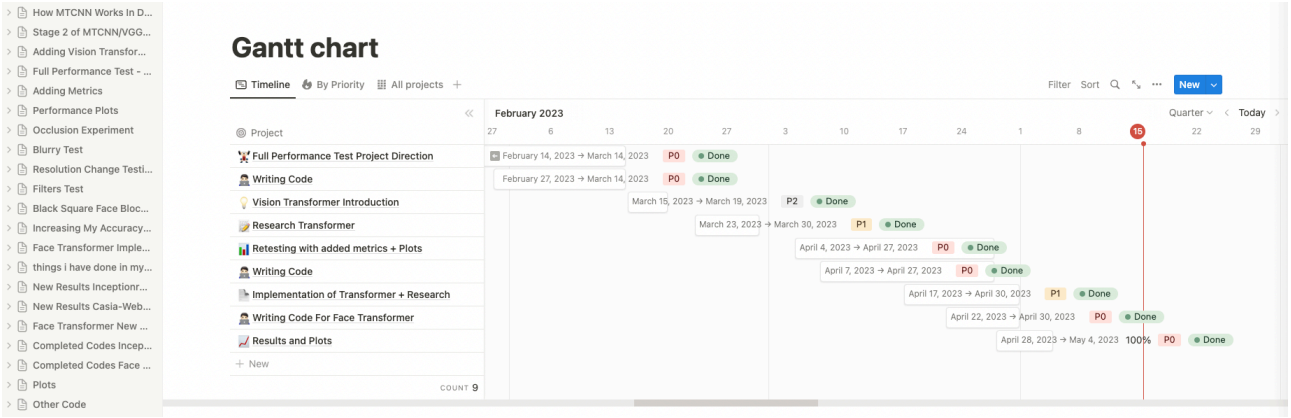


Figure 2: Notion page for this project including my Gantt chart to track progress

A link to all the research and progress on Notion is given in the references [Bauan, 2023].
Weekly meetings were taken with my supervisor to track progress and to make sure the project was moving ahead which was indispensable.

# 2: Literature Review

## 2.1: Introduction to Face Recognition

Face recognition focuses on identifying or verifying an individual's identity using their facial features. This technology has gained prominence due to its applications mainly in security and surveillance [Li et.al. 2011].

## 2.2: Key Challenges in Face Recognition Technology

Face recognition has several challenges, including occlusions(when various facial features are hidden) where Deng et al. (2021) highlighted the importance of occlusion as a benchmark for face recognition performance, which are common in real-world scenarios as they can partially or fully obscure facial features (Zhao et al., 2020). Other stress tests should be considered for face recognition models to assess their robustness under various conditions of the model such as: pose variations, illumination changes, beautification filters (Garcia et al., 2022), and facial expressions (A. Jain et al., 2011).

## 2.3: Pre-trained Face Recognition Models

Pre-trained models are trained on large datasets and often open sourced. The models used in this study: The Multi-task Cascaded Convolutional Networks (MTCNN) [Zhang et al., 2016], InceptionResNetV1 [Szegedy et al., 2017], and the new face transformer [Zhong et al. 2021].

### 2.3.1: MTCNN and InceptionResNetV1

MTCNN is a three-stage cascaded framework that detects facial regions and key points through a series of convolutional layers [Zhang et al., 2016]. InceptionResNetV1, a hybrid architecture derived from Inception-v4 and ResNet, is designed to provide highly accurate face recognition by combining the strengths of both architectures [Szegedy et al., 2017]. The pre-trained models used in this study are implementations of the facenet-pytorch Github repository[Timesler, 2019].

### 2.3.2: The New Face Transformer

The face transformer, proposed by Zhong et al. (2021), is a novel approach to face recognition that employs the self-attention mechanisms introduced in the Transformer architecture [Vaswani et al., 2017]. The Transformer architecture is designed for sequence-to-sequence tasks and is known for its ability to model long-range dependencies and complex relationships between input data. In face recognition tasks, the self-attention mechanism enables the model to focus on local facial features and their relationships, which may provide improved accuracy and robustness under varying conditions such as occlusion testing [Zhao et al., 2020]. The pre-trained face transformer model in this study is an implementation of the face transformer which can be found in the Github repository [Zhong, 2021].

## 2.4: Datasets used for pre-training in this study

Datasets play a crucial role in the performance of face recognition models.

The InceptionResNetV1 used in this study has been pre-trained on the VGGFace2 [Cao et al., 2018] and Casia-WebFace [Yi et al., 2014]. VGGFace2 is a large-scale dataset containing 3.31 million images of 9,131 subjects which means 362 images per subject. The Casia-WebFace contains 494,414 images of 10,575 subjects which means 47 images per subject. This difference gives the VGGFace2 7.7 times more diversity of images, making the model more robust for a wider range of inputs as hypothesised.

For pre-training the face transformer, the MS-Celeb-1M [Guo et al., 2016] was used in this study. This dataset contains around 10 million images of approximately 100,000 subjects from around the world.

### 2.5: Comparative Studies

Many comparative studies have been done to evaluate the performance of different face recognition models and machine learning architectures [Hameed et al., 2012]. For example, in the paper "DeepFace: Closing the Gap to Human-Level Performance in Face Verification" [Taigman et al. 2014], they compare the performance of different face recognition models, including their own DeepFace model, on the LFW dataset. These studies help researchers better understand the strengths and weaknesses of different methods and indicate that model performance varies depending on the architecture, training dataset, and evaluation conditions. The similarities between Taigman's study and our own are in testing the models on the LFW dataset for face verification. However, we differentiate our study by examining the performance of InceptionResNetV1 against the face transformer under a range of stress tests, including occlusion and image alterations using various metrics. We also compare the impact of VGGFace2 and Casia-WebFace datasets on the InceptionResNetV1 model's performance.

## 3: Methodology

### 3.1: Data Preparation

Figure 3 displays how the LFW dataset [Zhao et al., 2003], which contains 13,233 images with 5749 different people, is split.

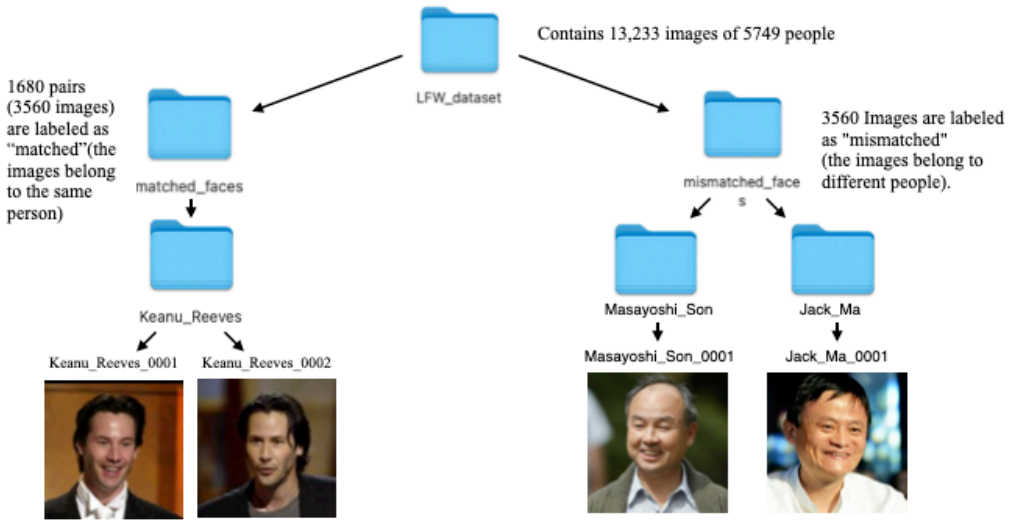


Figure 3: Data preparation to perform face verification for this study using the LFW dataset

In this study, the dataset is split into two categories: matching faces and mismatched faces. Matched faces are pairs of images for the same individual, while mismatched faces are pairs of images for two different individuals. The model has to verify whether the images in each pair represent the same person or are different individuals.

There are 5749 different people in the LFW dataset. 1680 are in the matched faces folder. The remaining 4069 are in the mismatched folder. However, to ensure controllability, 1680 pairs (3560 images of different people) were used for the mismatched folder. The remaining 509 faces of different people in the mismatched folder were not used.

This balanced approach allows for a fair comparison of this study so that the model's performance in verifying both matched and mismatched faces contain the same number of faces, providing valuable insights into the model's strengths and weaknesses in both tests.

### 3.1.1: Alterations made to the LFW dataset for the stress tests

Several alterations were made to the LFW dataset to test the models' performance under different conditions. These alterations are shown in Figure 4:


Figure 4: All Tests Done for Occlusion Testing

- Applying various levels of blurriness (intensity level from 1 to 5) to the images.
- Adding squares of various sizes in occlusion testing for hiding facial landmarks [Zhao et al., 2020]
- Converting the images to different resolutions (pixels): 48x48; 64x64; 96x96; 160x160.
- Applying different filters to the images which include grayscale, color tint, increasing brightness by 100%, and increasing contrast by 50% [Hameed et al., 2012].

### 3.1.2: Justifying Choices of Tests

Blurriness: Blurriness represents the loss of sharpness in an image, typically applied through filters like Gaussian or motion blur. This happens when there is frequent movement in an image or frame.
Occlusion: Testing whether the model can handle the loss of facial landmarks. This condition is commonly encountered in practical applications [Zhao et al., 2020].
Resolution: Resolution refers to the pixel count in an image, impacting the detail level which can involve resizing the image using interpolation methods. With different camera qualities, varying distances of faces in the frame, and image alterations, this becomes crucial in real world scenarios.
Filters: With the advent of social media and smartphones, the amount of image alterations using filters has drastically increased. Also, Hameed (2012) makes it clear why testing between different filters would be needed in order to perform face recognition adequately in real-world scenarios.
To evaluate the models' performance for real-world face recognition, it was logical to include the aforementioned stress tests.

## 3.2: Experimental Setup

We ensured the hardware and software conditions to test according to the scientific principles.
The hardware used was an M1 Mac Mini with an 8-core CPU and an 8-core GPU. All tests were run on the CPU, without any other heavy programs running simultaneously which ensured the CPU was close to equal in performance for each test. The system had 8GB of unified memory and 512GB of SSD storage. To replicate the experiments in this study, the GitHub repositories are available at [Cleary, 2023].
Regarding software, Python was used for writing the entire codebase because the libraries, repositories and the pre-trained models used in this study have also been written in python. Apart from any third-party libraries used, the only other code that was not written by us was from the pre-trained models (namely MTCNN, InceptionResNetV1 and Face transformer) which are located in their respective repositories [Timesler, 2019] and [Zhong, 2021].

## 3.3: Accounting for Anomalies

In order to adhere to scientific principles for analysing models with and without face detection algorithms, any undetected faces by the MTCNN would be given a random distance to control the evaluation between the InceptionResNetV1 and the face transformer(not using face detection).

## 3.4: MTCNN for Face Detection: Theory

MTCNN is a face detection algorithm that uses a cascaded structure of deep convolutional neural networks to achieve high accuracy and low computational time. [Zhang et al., 2016]. This section will describe the MTCNN algorithm in detail and the importance of facial landmarks generated by

MTCNN. This is to be used by the InceptionResNetV1 for extracting the embeddings. A visual representation of MTCNN is demonstrated in Figure 5.

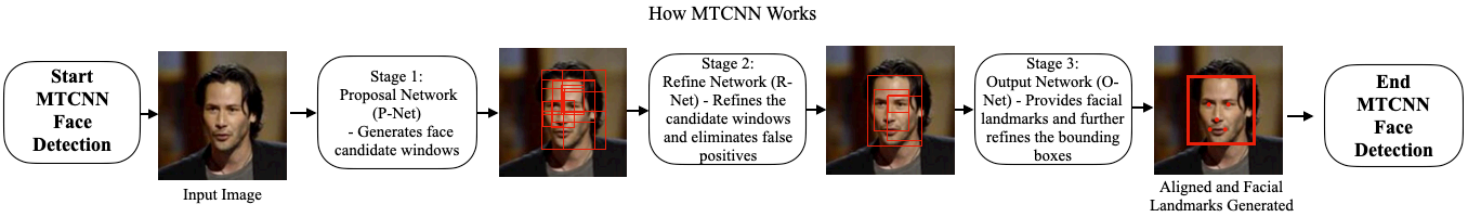


Figure 5: Step by Step guide on how MTCNN Works

**Stage 1 P-Net (Proposal Network):**

The P-Net is the first stage of the MTCNN algorithm and is responsible for the initial face detection. It takes an input image and creates an image pyramid to detect faces of different sizes [Zhang et al., 2016]. For each scaled image, a 12x12 kernel scans for faces by moving across the image with a stride of 2 pixels, reducing computation without significantly affecting accuracy [Zhang et al., 2016].

This process results in bounding boxes that are potential face regions with associated face confidence scores. The face classification loss function used for training is given by:

$$L_{\text{cls}} = -[y \cdot \log(p) + (1 - y) \cdot \log(1 - p)]$$

where $y$ is the ground truth label (1 for face, 0 for non-face) and $p$ is the predicted probability of a face.

**Stage 2 R-Net (Refine Network):**

The R-Net refines the bounding box coordinates, further eliminating false detections. R-Net takes the bounding boxes produced by P-Net as input and leverages its more complex architecture to improve the accuracy of face localisation [Zhang et al., 2016]. Non-Maximum Suppression (NMS) is applied after the R-Net stage to remove overlapping bounding boxes, leaving the most confident predictions [Zhang et al., 2016]. The bounding box regression loss function used for training is given by:

$$L_{\text{box}} = \sum (gt_i - \text{pred}_i)^2$$

where $gt_j$ is the ground truth offset value and $pred_j$ is the predicted offset value.

**Stage 3 O-Net (Output Network):**

The final stage, O-Net, further refines the bounding boxes and also generates facial landmark coordinates, such as the positions of the eyes, nose, and mouth [Zhang et al., 2016]. The O-Net produces three outputs: the probability of a face in the bounding box, the bounding box coordinates, and the facial landmark coordinates [Zhang et al., 2016]. These facial landmarks play a crucial role when InceptionResNetV1 extracts embeddings. The facial landmark localisation loss function used for training is given by:

$$L_{\text{landmark}} = \sum (gt_j - \text{pred}_j)^2$$

where $gt_j$ is the ground truth landmark coordinate and $pred_j$ is the predicted landmark coordinate. The total loss for each network is a weighted sum of the above loss functions:

$$L_{\text{total}} = L_{\text{cls}} + \lambda_1 \cdot L_{\text{box}} + \lambda_2 \cdot L_{\text{landmark}}$$

where $\lambda_1$ and $\lambda_2$ are hyperparameters that balance the contributions of each loss term.

**Techniques for Efficiency and Accuracy:**

- Image Pyramid: MTCNN processes the input image at multiple scales to detect faces of different sizes. This is achieved by creating an image pyramid, where each level is a scaled version of the input image.
- Sliding Window: For each scaled image, a sliding window with a fixed kernel size moves across the image with a stride of 2 pixels (this stride can be varied to suit computation requirements).
- Non-Maximum Suppression (NMS): After the R-Net and O-Net stages, NMS is applied to remove overlapping bounding boxes, retaining the most confident predictions. NMS is based on the Intersection over Union (IoU) metric:

$$\text{IoU} = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

If the $IoU$ between two bounding boxes exceeds a certain threshold, the one with a lower confidence score is suppressed.

By understanding the loss functions and techniques for efficiency and accuracy of the MTCNN, you can get a comprehensive view of the theory behind the face detection model.

### 3.4.1: Facial Landmarks and InceptionResNetV1 Embeddings

Facial landmarks generated by the MTCNN algorithm are crucial when InceptionResNetV1 extracts embeddings for face verification. These landmarks allow for accurate face alignment, which is a critical preprocessing step for deep face recognition systems [Zhang et al 2016]. Precise alignment ensures that the neural network can focus on facial features rather than variations caused by pose, expression, or illumination [Masi et al 2019].

When InceptionResNetV1 processes the face images, it extracts high-level features from the aligned faces, resulting in embeddings that are more robust [Schroff et al 2015]. By providing well-aligned input images, MTCNN significantly contributes to the performance of the InceptionResNetV1-based face recognition system.

## 3.5: InceptionResNetV1 for Face Recognition: Theory

InceptionResNetV1 is a powerful deep learning architecture that combines the strengths of the Inception architecture [Szegedy et al., 2016] and residual connections [He et al., 2016] to achieve state-of-the-art performance in various computer vision tasks, including face recognition. This model extracts facial features from images and generates compact vector embeddings that can be used for tasks such as face recognition, clustering, and classification [Schroff et al., 2015]. This section will present the theory behind the InceptionResNetV1 architecture.

**Inception Modules:**

The Inception modules form a core part of our network structure, operating by capturing and interpreting image features at different scales. This is accomplished through parallel branches, each employing convolutional filters of varying sizes (namely 1x1, 3x3, and 5x5). The small filters pick up fine details, and the larger filters see the bigger picture. After these filters have taken a look at the image, each one comes up with a sort of 'map' of what it has seen - these are the output feature maps. Each map highlights the features that its corresponding filter has picked up. The output feature stacks these feature maps together, allowing the network to learn complex and abstract features.

The convolutional operations in these modules can be represented mathematically as:

$$F_l = F_{l-1} + \text{Conv}(F_{l-1}, W_l) + b_l$$

where $F_1$ is the feature map at layer l, $F_{l-1}$ is the input feature map from the previous layer, $W_l$ and $b_l$ are the weight and bias parameters, and Conv represents the convolution operation.

**Residual Connections:**

Residual connections improve the training process and avoid degradation problems in deep networks. They allow the network to learn residual functions instead of the entire input-output mappings. Hence, simplifying the learning task.

A residual connection can be represented mathematically as:

$$F_l = \text{Conv}(F_{l-1}, W_l) + b_l$$

where the same notations from the inception modules are used.

**Global Average Pooling:**

After passing through the inception modules and residual connections, the feature maps are reduced in size using global average pooling. This operation reduces the spatial dimensions while retaining the high-level information captured by the feature maps. Mathematically, global average pooling can be expressed as:

$$P_i = \frac{1}{N} \sum F_i(x, y)$$

where $P_i$ is the pooled output for feature map i, $F_i(x, y)$ represents the feature map at spatial coordinates $(x, y)$, and N is the total number of spatial locations in the feature map.

**Embedding Generation:**

The output of the global average pooling layer is then passed through a fully connected layer, which generates a compact vector embedding representing the facial features of the input image. This operation can be represented mathematically as:

$$E = W \cdot P + b$$

where E is the vector embedding, W and b are the weight and bias parameters, and P is the pooled output from the global average pooling layer.

**Output Processing:**

The generated embeddings can be further processed, depending on the specific task at hand. The output is then used for face verification/recognition or classifications tasks. In our study the embeddings are compared using PyTorch as done in the facenet-pytorch repository [Timesler, 2019]. The schematic diagram of the InceptionResNetV1 is shown in Figure 20 in the appendix.

## 3.6 The Face Transformer for Face Recognition: Theory

The Face Transformer model recognises faces through the following steps as shown on Figure 6:

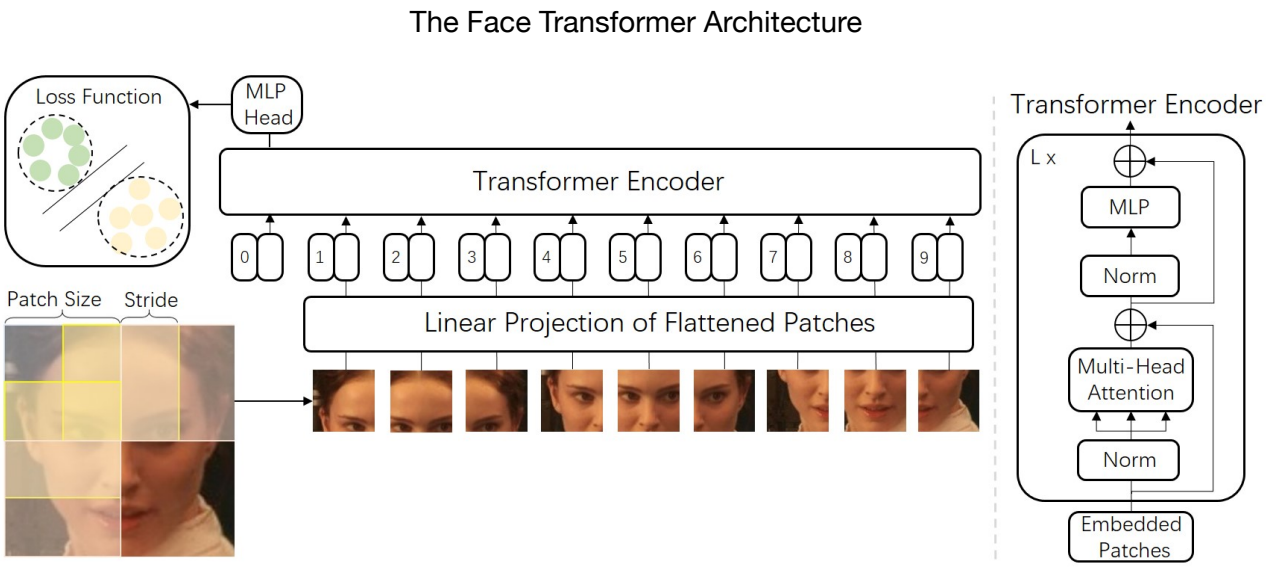


**Figure 6:** The overall architecture of the Face Transformer. Figure taken from [Zhong et al. 2021]

**Patch Extraction:**

The input face image is divided into multiple overlapping patches, capturing inter-patch information and local facial features. The patch extraction can be formulated mathematically as:

$$P_i = f(I, s, k, i)$$

where $P_i$ is the i-th patch, I is the input face image, s is the stride, k is the patch size, and f represents the patch extraction function.

**Patch Embeddings:**

Each patch is mapped to an embedding using a trainable linear projection, converting the patch into a fixed-size vector representation that can be processed by the Transformer model. This can be represented mathematically as:

$$E_{p_i} = W_p \cdot P_i + b_p$$

where $E_{pi}$ is the patch embedding, $W_p$ and $b_p$ are the weight and bias parameters, and $P_i$ is the i-th patch.

**Position Embeddings:**

Position embeddings are added to the patch embeddings to retain the positional information of each patch within the input image. This ensures that the model understands the spatial relationships between different patches. The addition of position embeddings can be expressed mathematically as:

$$E'_{p_i} = E_{p_i} + E_{\text{pos}_i}$$

where $E'_{pi}$is the updated patch embedding, $E_{pi}$ is the original patch embedding, and $E_{pos_i}$ is the position embedding for the i-th patch.

**Transformer Encoding:**

The concatenated patch and position embeddings are fed into the Transformer model, which consists of multi-headed self-attention and MLP blocks with LayerNormalisations and residual connections. The model processes the embeddings, focusing on important local features and their relationships within the face. The Transformer encoding can be represented mathematically using the following equation:

$$E_T = \text{Transformer}(E'_p)$$

where $E_T$ is the output of the Transformer model and $E'_p$ is the input patch embeddings with position information.

**Face Embeddings:**

The output of the Transformer model represents the face embeddings, which are unique vector representations of the input face images that can be used for face recognition tasks.

**Loss Function and Training:**

During training, the face embeddings are supervised using a softmax-based loss function that improves the model's discriminative ability. This loss function removes the bias term which transforms the output to the cos $\vartheta$ space, and then incorporates a large margin to enhance performance. The loss function can be expressed mathematically as:

$$L = -\log\left(\frac{\exp(s \cdot \cos(\theta_{y_i} + m))}{\sum \exp(s \cdot \cos(\theta_i))}\right)$$

where L is the loss, s is a scaling factor, $\theta_{yi}$ is the angle between the input face embedding and the correct class, m is the margin, and $\theta_i$ are the angles between the input face embedding and all classes.

**Face Recognition:**

Similar to the embedding comparison of the InceptionResNetV1, to recognise a face, the embeddings of the two faces it has to compare using the cosine distance metric. Faces with similar embedding values are considered to be the same person, while those with dissimilar embedding values are considered to be different individuals. The distance metric for face classification where it compares an input face to a face in the database can be represented mathematically as:

$$D = \text{distance}(E_{\text{input}}, E_{\text{database}})$$

where D is the distance metric, $E_{input}$ is the embedding of the input face, and $E_{database}$ are the embeddings of the reference faces in the database.

# 4: Evaluation

## 4.1 Using Embedding Extraction (Face Verification) Instead of Face Classification

Face recognition is an umbrella term that encompasses two main tasks: face verification is determining whether two facial images belong to the same person and face identification is matching a facial image to a database of known individuals to determine the identity of the person in the image. Both tasks involve analysing facial features to determine a match but used for different purposes and challenges.

Face identification can be further divided into two sub tasks: face classification and face retrieval. Face classification recognises and classifies specific individuals based on their facial features while face retrieval is searching through a database to find the closest facial images to a given image within the database.

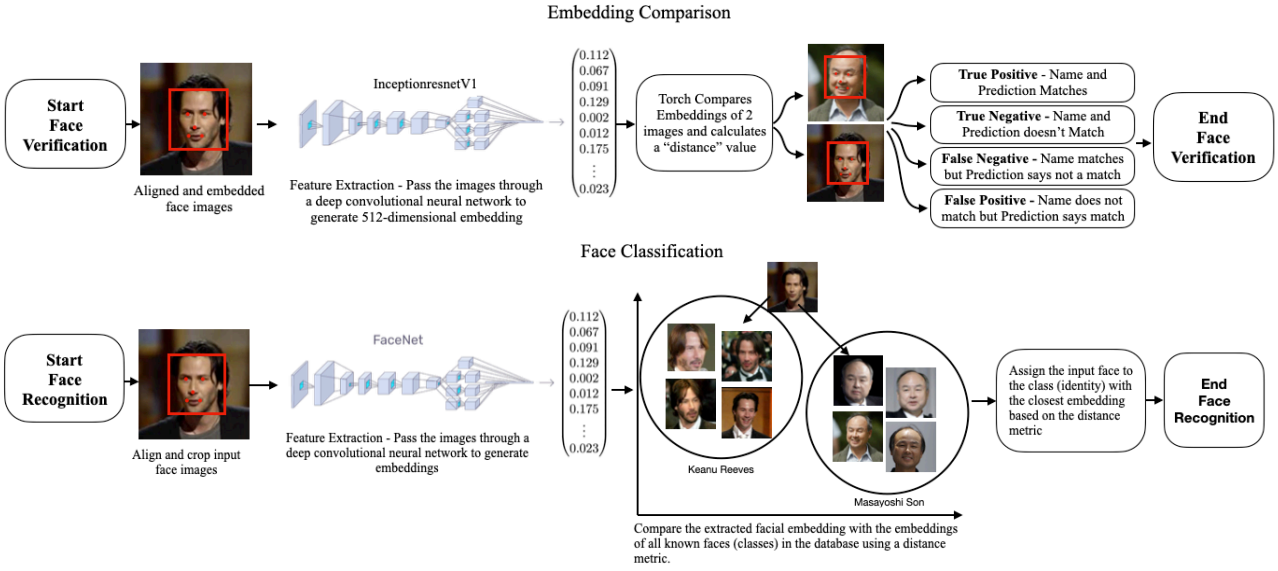


**Figure 7:** The difference between embedding comparison(Face verification), and face classification

In this study, we focus on extracting facial embeddings for face verification [Taigman et al., 2014] instead of directly obtaining classification results from the InceptionResNetV1 model [Szegedy et al., 2016]. The differences between the 2 methods are displayed in Figure 7. The primary motivation behind this decision is to ensure a fair comparison between the models and methods. To ensure this, the input for the InceptionResNetV1 and the face transformer have to both take embeddings as inputs [Parkhi et al., 2015].

By opting for verification over classification results, we may observe lower performance metrics in certain scenarios [Huang et al., 2007]. This is primarily because the classification output would typically include information regarding the specific identity of the subject, which can lead to higher accuracy in face recognition [Sun et al., 2014]. In contrast, embedding comparisons capture more general facial features and do not inherently contain identity-specific information.

Ultimately, the choice to use verification is justified in the context of our study, as our primary goal is to effectively compare the InceptionResNetV1 and the face transformer models.

## 4.2: Evaluation Metrics and their Importance

The evaluation metrics used in this research include accuracy, precision, recall, F1 score, Equal Error Rate (EER), optimal accuracy threshold, and computational time. These metrics are essential for understanding the performance of face recognition models and identifying the most suitable model for specific applications following the guidelines in the paper "A critical analysis of metrics used for measuring progress in artificial intelligence" [Blagec et al., 2020].

The base metrics are calculated based on the following: True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN), computational time, False Acceptance Rate and False Rejection Rate.

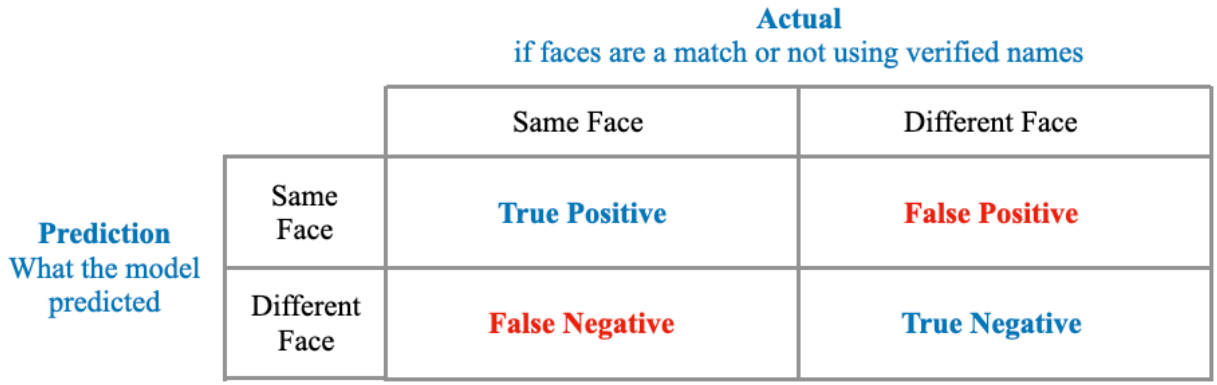


Figure 8: The difference between True Positive, True Negative, False Positives and False Negatives

**To explain each metric used in this study:**

True Positives, True Negatives, False Negatives, and False Positives (as shown in Figure 8): for example, if the model is presented with an image of Keanu Reeves and an image of Masayoshi Son, a True Negative would be for the model to predict they are different people. If the model predicts them to be the same person, then that is a False Positive.

Equal Error Rate (EER): This is the point where the False Acceptance Rate (FAR) and False Rejection Rate (FRR) are equal. Simply, EER is when the model has the same chance of letting in a stranger thinking they're the owner (FAR) as it does for locking out the real owner thinking they're a stranger (FRR). A lower EER indicates a better model performance.

Accuracy: This is the proportion of correct predictions (both true positives and true negatives) made by the model out of the total number of predictions. In face verification, accuracy reflects how well the model can correctly identify genuine pairs (individuals matched with their correct identity) and imposter pairs (individuals matched with a wrong identity). A higher accuracy indicates a better overall performance of the model. In terms of our base metrics:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: This is also known as positive predictive value which measures the proportion of true positives (correctly identified genuine pairs) among all the positive predictions made by the model. In simple terms, precision reflects how well the model can correctly identify genuine pairs without mistakenly accepting imposter pairs. A higher precision indicates a better model performance in distinguishing genuine from imposter pairs. In terms of our base metrics:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: This is also known as sensitivity or true positive rate which measures the proportion of true positives (correctly identified genuine pairs) among all the actual genuine pairs. In face verification, recall reflects the model's ability to identify genuine pairs without missing any genuine pairs. A higher recall indicates a better model performance in capturing all the genuine pairs. In terms of our base metrics:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score: In face verification, the F1 Score represents the balance between correctly identifying genuine pairs (precision) and capturing all genuine pairs (recall). A higher F1 Score indicates a better model performance, considering both precision and recall.

Optimal Threshold Accuracy: This metric is important because it helps determine the best threshold for making optimal face recognition predictions. A range of 5000 values was used between distances of 0.1 (predicting a matching pair) and 1.6 (predicting a mismatched pair). For each of the threshold values, the threshold that leads to the highest accuracy is calculated by reducing the False Acceptance Rate (FAR) and False Rejection Rate (FRR) as much as possible. An optimal threshold ensures that the model can best differentiate between matched and mismatched pairs.

Average Computation Time: A balance between speed and accuracy is ideal as a model with both high accuracy and low computation time is desirable for practical use.

Average Distance for Matched and Mismatched Faces: These metrics provide valuable insights into the model's confidence in making predictions. A low average distance for matched faces and a high average distance for mismatched faces indicates a more reliable and robust model which becomes apparent in figures 9-11.

### 4.2.1: Evaluation Approach

This study employs various visualisation techniques to analyse and compare the models' performance under different conditions. Percentage changes are used in the evaluation metrics, such as accuracy, recall, precision, F1 score, EER, and computational time, are plotted for each test which gives a comprehensive understanding of the models' behaviour under diverse challenges. Scatter plots are used to depict TP, TN, FP, and FN rates. Average distances for matched and mismatched faces are also used to offer a deeper insight into the models' strengths and weaknesses.

# 5: Results

## 5.1: Performance of the MTCNN and InceptionResNetV1 models

In this section, we analyse the performance of the MTCNN and InceptionResNetV1 models in terms of face detection and face recognition accuracy on the LFW dataset. Table 1 compares the MTCNN model with other face detection methods in terms of recall, precision, F1-score, and speed (FPS):

| Method | Recall | Precision | F1-Score | Speed (FPS) | Source |
|---|---|---|---|---|---|
| **MTCNN** | 0.97 | 0.98 | 0.975 | 20-25 | [Zhang et al., 2016] |
| **Cascade CNN** | 0.94 | 0.97 | 0.955 | 15-20 | [Li et al., 2016] |
| **Faster R-CNN** | 0.91 | 0.95 | 0.93 | 5-10 | [Ren et al., 2017] |
| **Multi-task Cascaded** | 0.90 | 0.94 | 0.92 | 10-15 | [Yang et al., 2016] |

**Table 1:** Performance of different Face Detection models

As seen in Table 1, the MTCNN model outperforms the other methods, achieving the highest F1-score and maintaining a relatively fast speed.

For face recognition, we evaluate the InceptionResNetV1 model, which serves as the base architecture for the FaceNet model. Table 2 compares the classification accuracy of the InceptionResNetV1-based FaceNet model with other popular face recognition models on the LFW dataset:

| Model | LFW Classification Accuracy (%) | Source |
|---|---|---|
| **FaceNet (Using Inception-resnetV1 as its base architecture)** | 99.63 | Schroff et al., 2015 |
| **Center Loss** | 99.28 | Wen et al., 2016 |
| **DeepFace** | 97.35 | Taigman et al., 2014 |
| **VGG-Face** | 98.95 | Parkhi et al., 2015 |

**Table 2:** Performance of different Recognition (Using Face Classification) models

The InceptionResNetV1-based FaceNet model achieves the highest classification accuracy among the models compared, demonstrating the effectiveness of the chosen combination of MTCNN and InceptionResNetV1 in the face recognition pipeline. This strong performance justifies the choice of these models for the face recognition tasks in this study.

## 5.2 Explanation of the results attained in this study:

The standard performance of each model when performing face verification on the LFW dataset is shown in Table 3.

| Model | Accuracy | Precision | Recall | F1 Score | EER | Average Computation Time (seconds) |
|---|---|---|---|---|---|---|
| **MTCNN/ InceptionresnetV1 (VGGFace2)** | 0.94 | 0.97 | 0.91 | 0.94 | 0.068 | 0.079 |
| **MTCNN/ InceptionresnetV1 (Casia-Webface)** | 0.92 | 0.94 | 0.90 | 0.92 | 0.082 | 0.085 |
| **Face Transformer (MS-Celeb-1M)** | 0.81 | 0.84 | 0.76 | 0.80 | 0.20 | 0.61 |

**Table 3:** Standard (112x112) Performance of the models that is being compared in this study

It is essential to note that the "standard performance" is where all images have to be resized to a 112x112 resolution by the MTCNN or face transformer. This will serve as the baseline for comparing all other tests. This approach is taken to assess the models' robustness against altered images. This comparison will be crucial because the percentage change is relative to the standard. This resizing is necessary because the face transformer has been trained using 112x112 sized images and cannot process different sizes. To ensure controllability in all of the tests, our standard models resized the images to this size.

Another example to maintain standardisation across all the tests is the optimal threshold accuracy. The optimal threshold accuracy must be determined from the standard model, only then are the models tested on all the other stress tests. This consideration is critical because optimal thresholds can only be determined with a large enough dataset, but in real-world scenarios, this is not possible. Therefore, only one optimal threshold accuracy is determined for the entire model. It is also important to remember that the threshold can be tailored to perform optimally for precision, F1 score, or any other metric. However in this study, we have chosen to achieve the highest accuracy for the models.

## 5.3: Performance of the MTCNN and InceptionResNetV1 (VGGFace2) Model in Face Verification

The plots presented in this study allowed us to identify the underlying weaknesses in each model and where improvements can be made. The standard InceptionResNetV1 (VGGFace2) model performed remarkably well, as shown in Table 3 and Figure 9. It is able to clearly differentiate between matched face pairs and mismatched face pairs. The average matching distance is 0.77 and the average mismatched distance is 1.342 while the EER value is 0.068. The EER value is so low because of the model's ability to almost equally reduce false positive and false negative results.

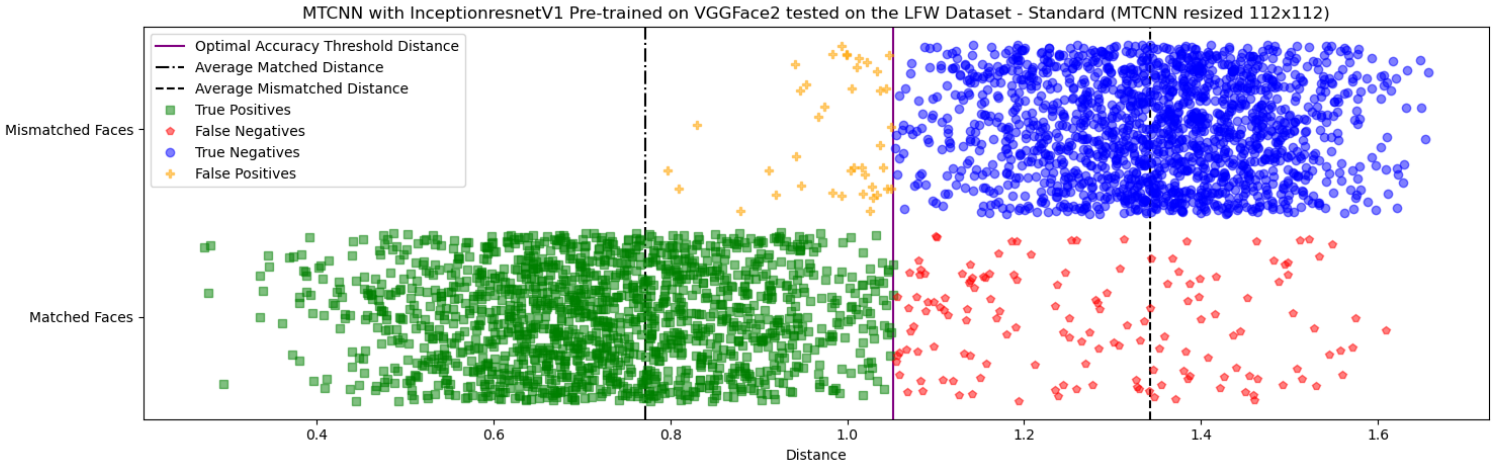


**Figure 9:** Plot illustrating TP, TN, FP, and FN for InceptionResNetV1 trained on VGGFace2

Figures 12-15 display the results of each stress test on the InceptionResNetV1(VGGFace2). It is evident that the performance significantly decreases for: blurry intensity 4 and above, resolutions of 48x48, and occlusion square sizes of 15% or higher where facial landmarks are likely to be lost. However, the grayscale filter does not appear to affect the results compared to other filters that negatively impact the model's performance. This is understandable, unlike the other filters because the grayscale filter mainly removes color information from the images, while still preserving the contrast, edges, and texture information, which are crucial for identifying facial features and landmarks.

## 5.4: Performance of the MTCNN and InceptionResNetV1 (Casia-Webface) Model in Face Verification

Figure 10 demonstrates why the standard model for the Casia-Webface performed slightly worse than the VGGFace2, even though they are the same models but pre-trained on different datasets. The answer lies in the 'confidence' the models have in their distance scoring. The average matched face distance for the VGGFace2 is 0.78 while the Casia-Webface is 0.82. This difference is because of the variation in the images within the datasets as stated in hypothesis 2.

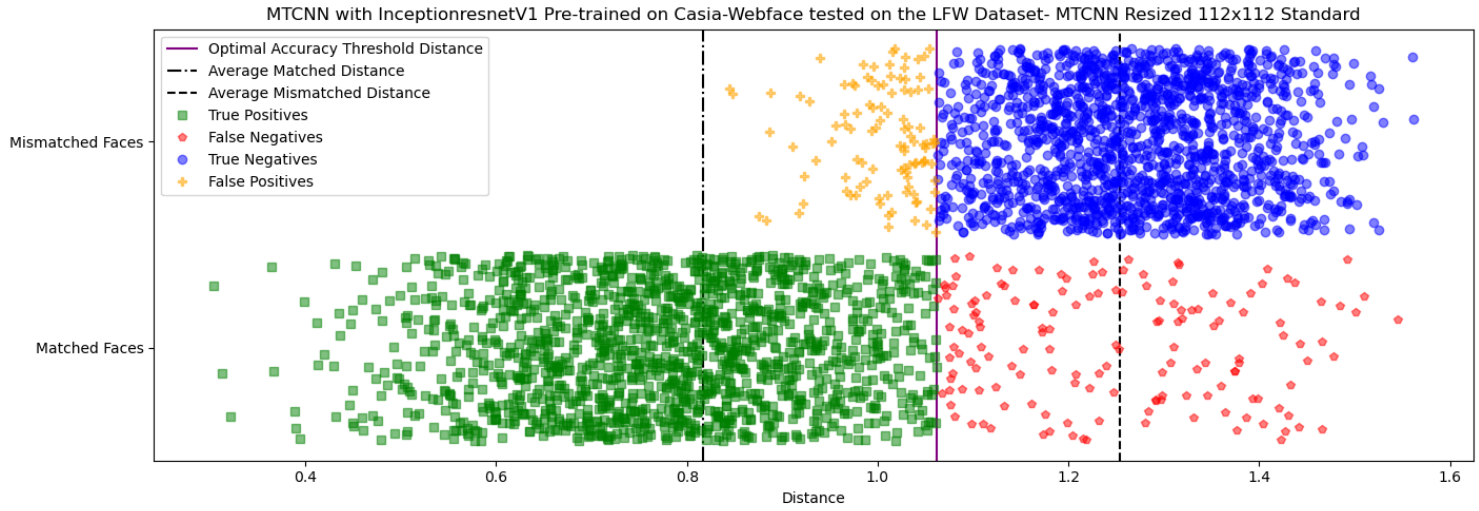


**Figure 10:** Plot illustrating TP, TN, FP, and FN for InceptionResNetV1 trained on Casia-Webface

In terms of stress tests, the Casia-Webface is roughly equally robust to the VGGFace2 model. However, the Casia-Webface model percentage change falls more significantly for metrics like accuracy and precision, especially when the blurry intensity is at 4 or higher.

## 5.5: Performance of the Face Transformer Model for Face Verification

The face transformer performed quite poorly compared to the InceptionResNetV1. This is in contrast to the claim in the Zhong et al. (2021) paper that the model performs similarly to equivalent CNN models of similar-sized datasets. Nevertheless, the face transformer is functional and can clearly differentiate between matched and mismatched pairs of faces. The optimal threshold accuracy was calculated to be 1.044, and it is also apparent that the distance values often extend far beyond those of the InceptionResNetV1 close to 3.2 for some mismatched faces. However, this is not necessarily a negative aspect because it has a strong confidence in some mismatched faces.


**Figure 11**: Plot illustrating TP, TN, FP and FN for the FaceTransformer trained on MS-Celeb-1M

A low average distance for matched faces and a high average distance for mismatched faces indicates a more reliable and robust model. Therefore, the reason for the poor results shown in Table 3 lie in the model's confidence in the distance values it assigns to the pairs shown in Figure 11. For example, the average distance of the matched pairs is 0.865 meaning the majority of the data falls closer to the threshold distance compared to the InceptionResNetV1 models. The face transformer also performed poorly in filter tests, such as: the color tint filter, an increase in contrast and brightness, change in blurriness, or decrease in the resolution of the images. However, as predicted in hypothesis 3, the face transformer does much better in the square occlusion tests compared to the InceptionResNetV1 models with an increase of computational time as a consequence.

## 5.6: Comparison of Results for All Models and Methods Tested

Figures 12-15 display the performance metrics for all models across various tests.

It is evident that, for all of the tests aside from the square occlusion tests, the InceptionResNetV1 pre-trained on the VGGFace2 dataset performs best. This demonstrates the model's ability to accurately identify pairs of images, which is challenging.

In terms of the face transformer's results, the model required significant computational time to run all tests, particularly for the square occlusion tests where it performed the best out of the three models. However, the face transformer model exhibited signs of overfitting and poor generalisation. This became clear with the following details: as illustrated in Figure 11 — here, the model does differentiate reasonably well between matched and mismatched faces.

However, for the blurry and resolution tests, as precision decreases for the face transformer, recall increases [Figure 18]. This maintains the EER and minimally impacts accuracy, emphasising the importance of multiple metrics for assessing model performance and tracking additional parameters. If only accuracy is used as a metric to track performance, this flaw in the face transformer model would not have been identified. Additionally, as seen in Figure 19, the average matching and mismatching pair distances both decrease as blurriness intensity increases. This highlights how the face transformer shifts the entire distance scores more towards 0 [Figure 20]. Here, the face transformer almost entirely avoids false negatives, but the number of false positives significantly increases. A similar phenomenon occurs with resolution changes for the face transformer. However, this is not the case for the InceptionResNetV1 as shown in Figure 17 where the mismatching distance and the matching distances become closer in value when it is faced with difficult stress tests such as the 20% square occlusion test. Here, the InceptionResNetV1 considers images it finds hard to detect (Figure 16) by the MTCNN to likely be a mismatching pair because of the lost facial landmarks generated by the MTCNN.

Furthermore, the face transformer performs worse than the other models when filters are applied, with the exception of the grayscale filter, where all models exhibit robustness and faster computational times [Figure 14].

As hypothesised, the face transformer performs remarkably well when the black square occlusion is applied to the LFW images, despite the sharp decline in performance for the InceptionResNetV1 models as shown in Figure 15. For the face transformer, there is a slight decrease in precision and a slight increase in recall for the 10%-20% square sizes. This is also accompanied by a significant spike in computational time.

The InceptionResNetV1 as well as the MTCNN struggle to recognise most images [Figure 16], almost coming to the point of guessing when the square occlusion size is at 20%. However, in contrast to the InceptionResNet models, the transformer performs incredibly well [Figure 15 & 21] in the occlusion tests.

## 5.7 Investigating the Performance of the Face Transformer on Blurry and Low Resolution Images

In line with the research objectives and hypotheses, this section explores the performance of the face transformer model in face recognition tasks, particularly under challenging conditions like blurry and low-resolution images. By analysing the model's behaviour and performance, we aim to uncover potential issues and suggest improvements for enhancing the reliability and effectiveness of face recognition technology.

**Hyperparameters and Model Architecture**

The face transformer model's performance is influenced by the choice of hyperparameters and architecture. In our experiments, we considered both the 'VIT' and 'VITs' architectures and tested various hyperparameter configurations to determine their suitability for our task [Dosovitskiy et al., 2021].

**Data Preprocessing and Augmentation**

Proper data preprocessing, including resizing and normalisation, is crucial for achieving optimal model performance. Data augmentation can further enhance generalisation by introducing variability in the training data, particularly for blurry and low-resolution images [Shorten & Khoshgoftaar, 2019].

**Loss Function and Regularisation**

The CosFace loss function was employed in our experiments. While it offers certain advantages, it might not be optimal for our specific task [Wang et al., 2018]. Other loss functions, such as ArcFace or SphereFace, could potentially improve performance on blurry and low-resolution images [Deng et al., 2019; Liu et al., 2017]. Regularisation techniques, like dropout, weight decay, and early stopping, can further enhance the model's generalisation capabilities.

**Potential Improvements and Experiments**

Based on my analysis, we propose the following improvements and experiments to address the issues faced by the face transformer model in handling blurry and low-resolution images:

- Augment the training data with more low-resolution and blurry images
- Fine-tune the model on low-resolution or blurry images
- Explore other loss functions such as ArcFace or SphereFace
- Experiment with different model architectures or hyperparameter configurations to improve the model's robustness
- Incorporate attention mechanisms to focus on the most discriminative facial features, especially in challenging conditions

# 6: Discussion:

## 6.1: Interpretation of Results

In light of the research objectives and hypotheses, the InceptionResNetV1 model, particularly when trained on the VGGFace2 dataset, demonstrated excellent performance and robustness to various stress tests, such as blurriness and filters [Cao et al., 2018]. This finding supports Hypothesis 1, indicating that pre-trained models like InceptionResNetV1 offer superior face recognition performance due to their established architectures and training on extensive datasets [Schroff, Kalenichenko, & Philbin, 2015].

Our results also revealed that the choice of dataset does impact performance, even when using the same architecture. This observation was evident in the differences between the models trained on VGGFace2 and Casia-WebFace datasets [Cao et al., 2018; Yi et al., 2014], thereby supporting Hypothesis 2. This finding highlights the importance of data distribution, sample size, and representation of facial variations within each dataset in determining the model's performance and robustness to occlusions and image alterations [Sun et al., 2014].

Regarding the face transformer, although it is still in its early stages and the first paper to introduce the transformer architecture to face recognition, the model shows potential in addressing the square occlusion tests [Dosovitskiy et al., 2020]. This observation supports Hypothesis 3, suggesting that the face transformer, with its patch-based approach that focuses on local features, could demonstrate improved robustness in occlusion tests where facial landmarks are hidden compared to traditional models [Vaswani et al., 2017]. However, there is still significant work needed to enhance other metrics and reduce computational time especially when dealing with high levels of occlusion [Figure 15].

A limitation of the pre-trained face transformer model presented in the paper is that it has not generalised well enough to verify matching faces when there are changes in blurriness and resolution. This issue presents a significant challenge for the model's overall performance and usability. The need for further research and development to improve generalisation capabilities are required in order to enhance face recognition accuracy across a wider range of real-world conditions [LeCun, Bengio, & Hinton, 2015].

## 6.2: Comparison of the Results with Previous Research

In comparing our results to previous research, several key differences and similarities emerge. Overall, our findings align with the literature, showing that pre-trained models such as InceptionResNetV1 demonstrate strong performance in face recognition tasks (Schroff, Kalenichenko, & Philbin, 2015). Our research also supports the notion that the quality of the dataset used for training significantly impacts model performance, as evidenced by the differences between the VGGFace2 and Casia-WebFace models [Cao, Shen, Xie, Parkhi, & Zisserman, 2018; Yi, Lei, Liao, & Li, 2014].

However, our study differs from previous research in its focus on the face transformer's performance in various stress tests.

Firstly, while the transformer has shown promise in early face recognition research [Zhong et al. 2021], our study provides a more in-depth evaluation of its robustness to occlusions and image alterations. We do this by comparing the face transformer to the InceptionResNetV1 under various metrics and extensively visualising the results.

Secondly, our research builds upon previous studies by examining the impact of diverse stress tests on face recognition models providing a more comprehensive understanding of model performances under challenging conditions such as blur, resolution and filters tests.

Thirdly, our study uses multiple metrics, such as precision and recall, to show the flaws in the face transformer model. If we had not used multiple metrics and plots our results would have been flawed as demonstrated by Zhong (2021), therefore emphasising the importance of transparency and repeatability

In conclusion, our study contributes to the existing body of research on face recognition models by providing a comprehensive comparison of pre-trained models and the face transformer under various tests. Also, our findings support the importance of dataset quality, model architecture, and training procedures in determining face recognition performance. Finally, we offer insights into potential areas for improvement in the face transformer's design.

## 6.3: Limitations of the Study and Areas for Future Research

Our study, like any research endeavour, has its limitations. One notable limitation is the lack of in-depth analysis for the face transformer's code in the Git Repository [Zhong, 2021], which could have provided a better understanding on the specific issues affecting its performance in various tests [Zhong et al. 2021]. Future research could delve deeper into the face transformer's architecture and training procedures to identify areas for improvement and optimise its performance under different conditions.

Another limitation of our study is the absence of a face transformer model trained on the VGGFace2 dataset and under the same conditions as the best-performing InceptionResNetV1 model [Cao et al., 2018]. This would have allowed for a more accurate comparison between the models and provided further insights into the potential benefits of using different datasets for training the face transformer. Future research should consider training the face transformer on multiple datasets and evaluating its performance under comparable conditions to other pre-trained models.

Furthermore, our study relied on the LFW dataset for testing [Huang et al., 2007] which may not fully represent the range of real-world conditions that face recognition models might encounter. Future research could include testing with multiple datasets, such as IARPA Janus Benchmark A (IJB-A) [Klare et al., 2015], to determine whether the performance of the models vary across different data sources, and to provide a more comprehensive understanding of their robustness to diverse conditions.

In our opinion, the biggest contribution that can be done working on top of our findings from this study is to retrain a face transformer model on the VGGFace2 dataset using the vision transformer from the paper by Dosovitskiy (2021). The VGGFace2 training dataset could be augmented by adding blurring, resolution changes and filters for better generalisation. Lastly, the model should be tested with metrics included in this study to avoid missing weaknesses in the model as done by Zhong (2021).

By addressing these limitations and incorporating these suggestions, future research can contribute to a more in-depth understanding of face recognition models' performance and inform the development of more robust and accurate models for various real-world applications.

# 7: Conclusion

## 7.1: Summary of the research findings

This study investigated the performance of various face recognition models, including InceptionResNetV1 and the face transformer, under different occlusion tests and image alterations. The results revealed that the InceptionResNetV1 model, especially when trained on the VGGFace2 dataset, demonstrated exceptional performance and robustness across various stress tests. The choice of dataset significantly influenced model performance, with VGGFace2 outperforming Casia-WebFace. The face transformer, while still in its early stages, showed potential in addressing specific occlusion challenges, such as square occlusions.

### 7.2: Implications of the research for face recognition technology

The findings of this study highlight the importance of selecting appropriate datasets and fine-tuning pre-trained models to optimise face recognition performance under diverse conditions. This needs to be coupled with a good set of metrics for determining the validity of a good model. Otherwise, there is a risk of releasing models that do not generalise well to edge cases.

### 7.3: Future directions for research

Future research should focus on addressing the limitations identified in this study, including exploring the face transformer model's code to pinpoint areas for improvement. The face transformer can also be trained on the VGGFace2 dataset under the same conditions as the InceptionResNetV1 model. Testing models with additional datasets beyond LFW would also provide a more comprehensive understanding of their performance. Furthermore, future research should continue to explore the potential of emerging face recognition models such as the face transformer.

## References (APA style):


1. Rougier, N. P., Hinsen, K., Alexandre, F., Arildsen, T., Barba, L. A., Benureau, F. C. Y., Brown, C. T., de Buyl, P., Caglayan, O., Davison, A. P., Delsuc, M. A., Detorakis, G., Diem, A. K., Drix, D., Enel, P., Girard, B., Guest, O., Hall, M. G., Henriques, R. N., . . . Zito, T. (2017). Sustainable computational science: the ReScience initiative. PeerJ Computer Science, 3, e142. https://doi.org/10.7717/peerj-cs.142
2. Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). VGGFace2: A dataset for recognising faces across pose and age. International Journal of Computer Vision, 126(2-4), 902-919.
3. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2021). When Face Recognition Meets Occlusion: A New Benchmark. arXiv preprint arXiv:2103.02805.
4. Garcia, S., Levada, A. L. M., & de Rezende Rocha, A. (2022). On the effect of selfie beautification filters on face detection and recognition. Pattern Recognition Letters, 156, 90-98.
5. Hameed, K. A., Ismael, A. M., & Al-Temeemy, S. M. (2012). A comparative study on face recognition techniques and neural network. arXiv preprint arXiv:1210.1916.
6. Masi, I., Tran, A. T., Hassner, T., Sahin, G. G., & Medioni, G. (2019). Face recognition under varying blur, illumination and expression in an unconstrained environment. arXiv preprint arXiv:1902.10885.
7. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). FaceNet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 815-823).
8. Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017). Inception-v4, Inception-ResNet and the impact of residual connections on learning. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 31, No. 1).

9. Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In European conference on computer vision (pp. 499-515). Springer, Cham.
10. Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Learning face representation from scratch. arXiv preprint arXiv:1411.7923.
11. Zhang, K., Zhang,Z., Li, Z., & Qiao, Y. (2016). Joint face detection and alignment using multi-task cascaded convolutional networks. IEEE Signal Processing Letters, 23(10), 1499-1503.
11. Zhao, W., Chellappa, R., Phillips, P. J., & Rosenfeld, A. (2003). Face recognition: A literature survey. ACM Computing Surveys (CSUR), 35(4), 399-458.
12. Zhao, S., Gao, J., Han, R., & Shan, S. (2020). A survey of face recognition techniques under occlusion. arXiv preprint arXiv:2006.11366.
13. Zhong, Y., Deng, W., Hu, J., & Sun, J. (2021). Face Transformer for Recognition. arXiv preprint arXiv:2103.14803.
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).
15. Blagec, K., Dorffner, G., Moradi, M., & Samwald, M. (2020). A Critical Analysis of Metrics Used for Measuring Progress in Artificial Intelligence. arXiv preprint arXiv:2008.02577v2. https://arxiv.org/abs/2008.02577
16. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Uszkoreit, J. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations.
17. Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on Image Data Augmentation for Deep Learning. Journal of Big Data, 6(1), 60.
18. Wang, H., Wang, Y., Zhou, Z., Ji, X., Li, Z., Gong, D., ... & Liu, W. (2018). CosFace: Large Margin Cosine Loss for Deep Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
19. Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
20. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). SphereFace: Deep Hypersphere Embedding for Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
21. Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. University of Massachusetts, Amherst, Technical Report 07-49.
22. Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active Appearance Models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(6), 681-685.
23. Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. *Proceedings of the British Machine Vision Conference (BMVC)*, September 2015. DOI: 10.5244/C.29.41

24. Sun, Y., Wang, X., & Tang, X. (2014). Deep Learning Face Representation from Predicting 10,000 Classes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. DOI: 10.1109/CVPR.2014.241
25. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, Xiong, H., & Tang, X. (2018). Residual Attention Network for Image Classification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2017. DOI: 10.1109/CVPR.2017.439
26. Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). DeepFace: Closing the Gap to Human-Level Performance in Face Verification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. DOI: 10.1109/CVPR.2014.220
27. Li, Y., Sun, B., Wu, T., & Wang, Y. (2016). Face Detection with End-to-End Integration of a ConvNet and a 3D Model. arXiv:1606.00850.
28. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(6), 1137-1149.
29. Yang, T., Luo, P., Loy, C. C., & Tang, X. (2016). WIDER FACE: A Face Detection Benchmark. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
30. Guo, Y., Zhang, L., Hu, Y., He, X., & Gao, J. (2016). MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In Proceedings of the 14th European Conference on Computer Vision (ECCV).
31. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
32. Wang, Z., Zhao, J., Wang, Y., Zhang, J., Li, X., & Peng, C. (2021). Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. arXiv preprint arXiv:2102.12122.
33. Klare, B. F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., & Burge, M. (2015). Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1931-1939.
34. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778.
35. Timesler. (2019). Facenet-pytorch. GitHub, from https://github.com/timesler/facenet-pytorch
36. Zhongyy. (2021). Face-Transformer. GitHub, from https://github.com/zhongyy/Face-Transformer
37. Cleary, T. (2023). Repositories. GitHub, from https://github.com/DrThomasCleary?tab=repositories
38. Bauan, R.(2023). Independent Project. Notion. Retrieved from: https://secretive-passbook-b9c.notion.site/bd3d7c7d01c74606961748109d321f34
39. Li, S. Z., & Jain, A. K. (2011). Handbook of Face Recognition. Springer.

# Appendix

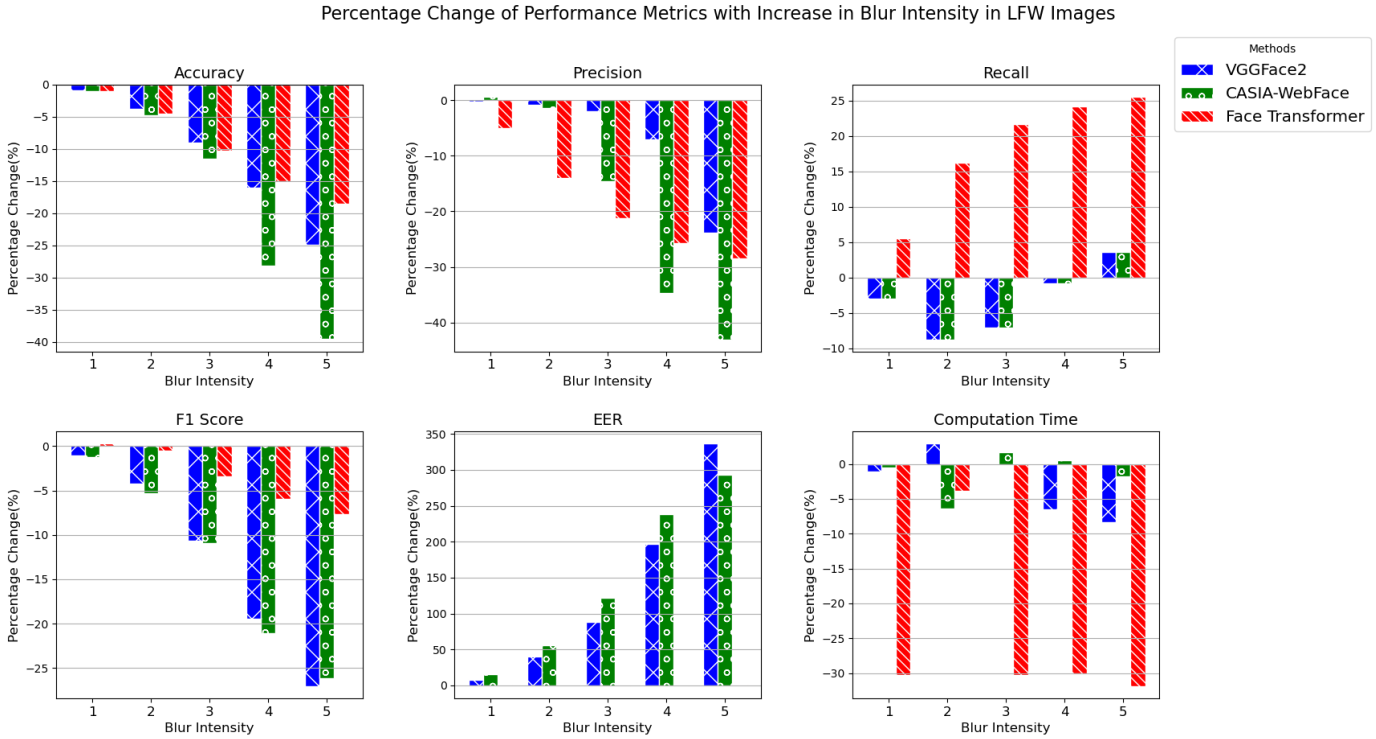


**Figure 12:** Percentage Change of all metrics in each model with respect to changes in Blur Intensity

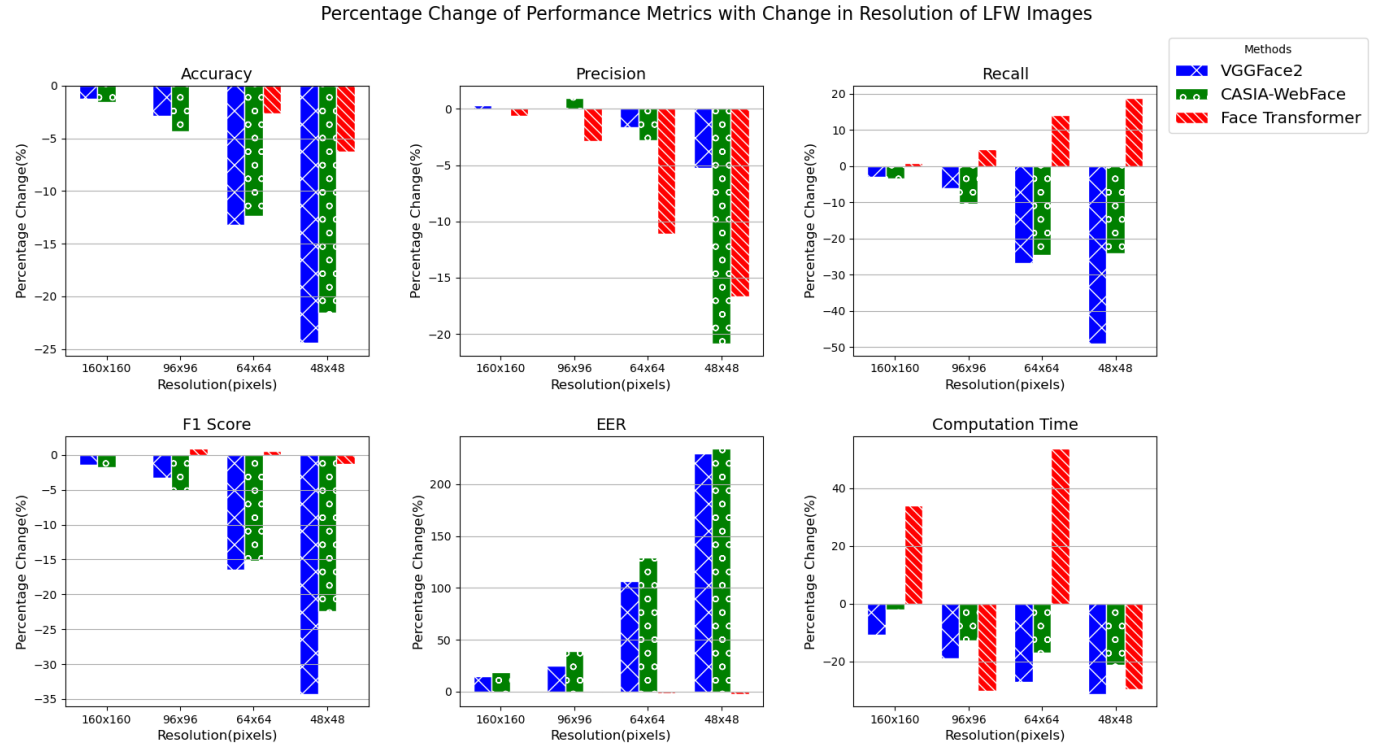


**Figure 13:** Percentage Change of all metrics in each model with respect to changes in Resolution

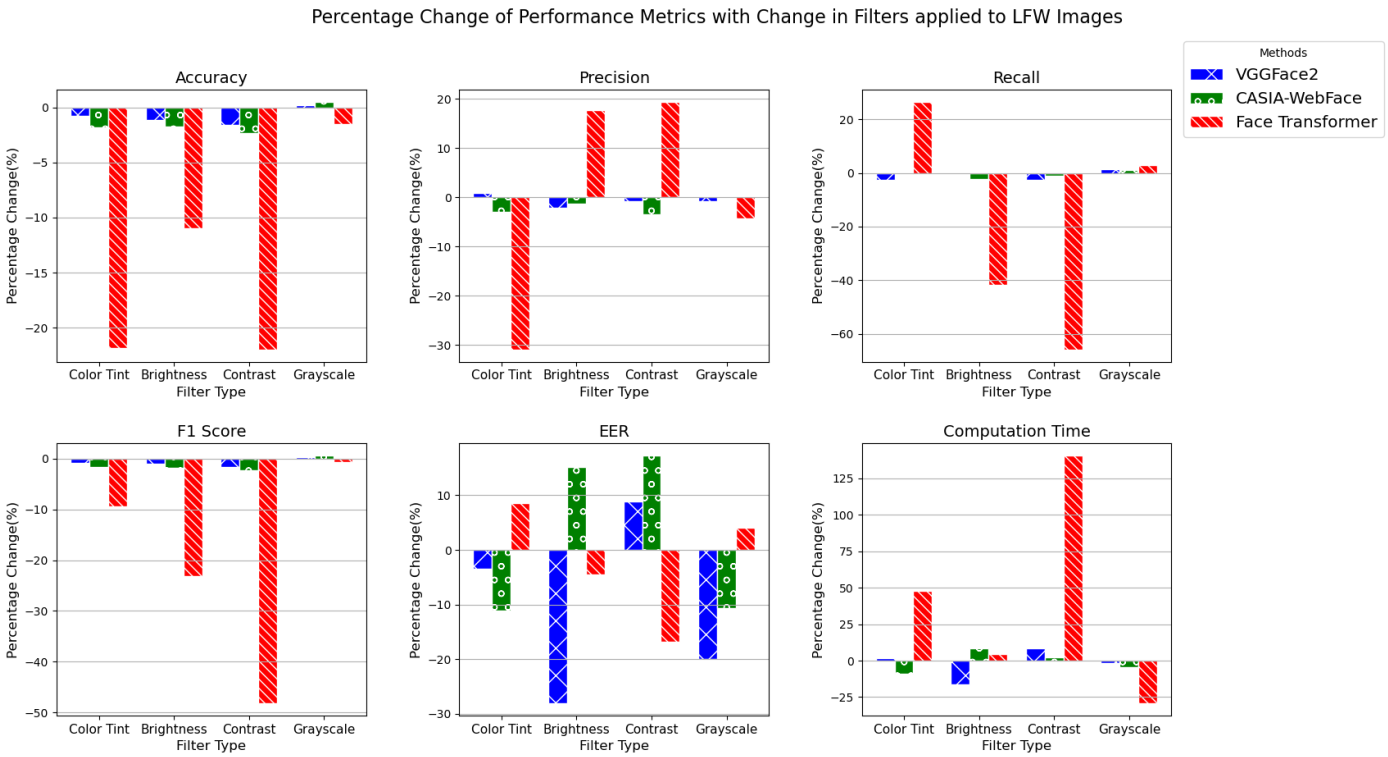


**Figure 14:** Percentage Change of all metrics in each model when filters are applied

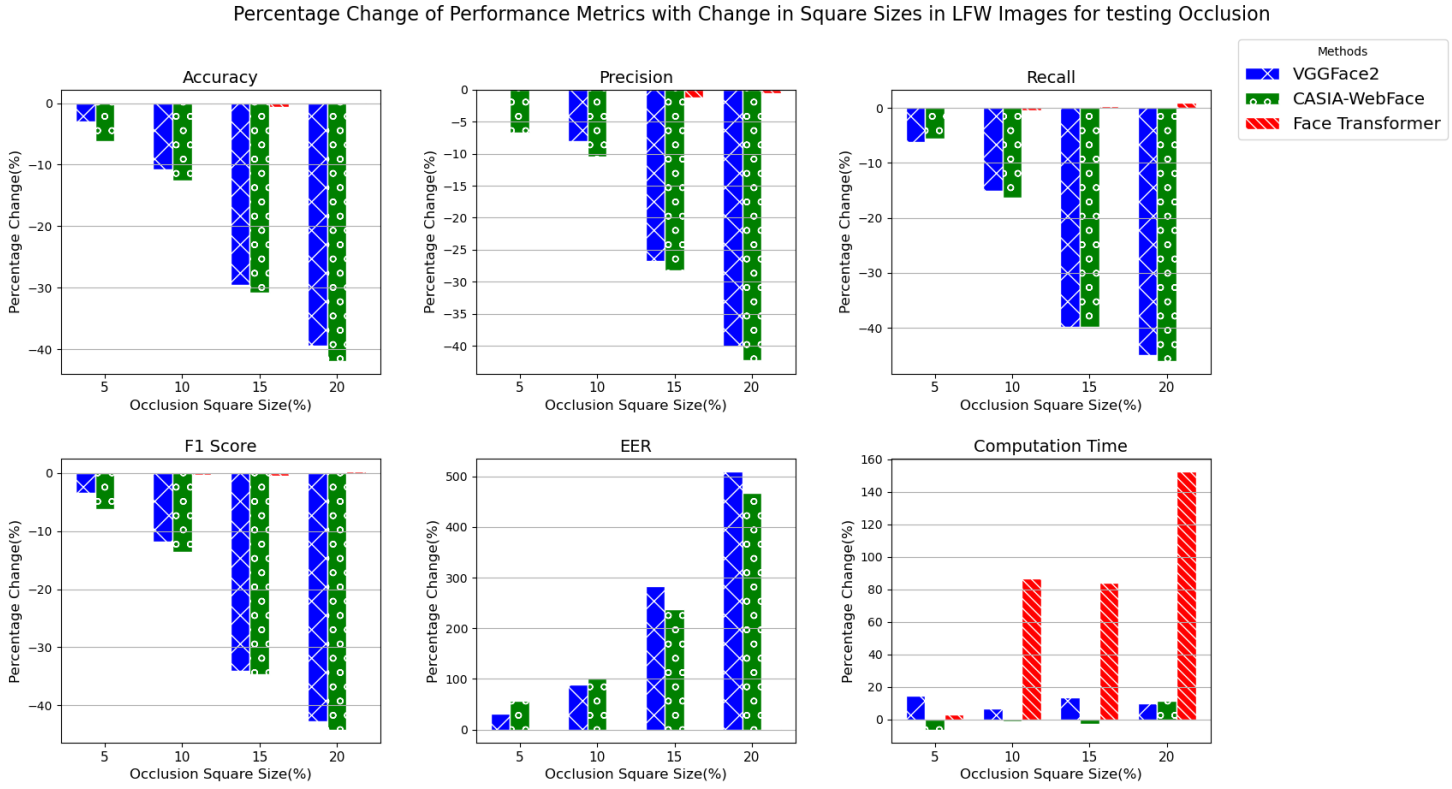


**Figure 15:** Percentage Change of all metrics in each model with increase of square size for occlusion testing

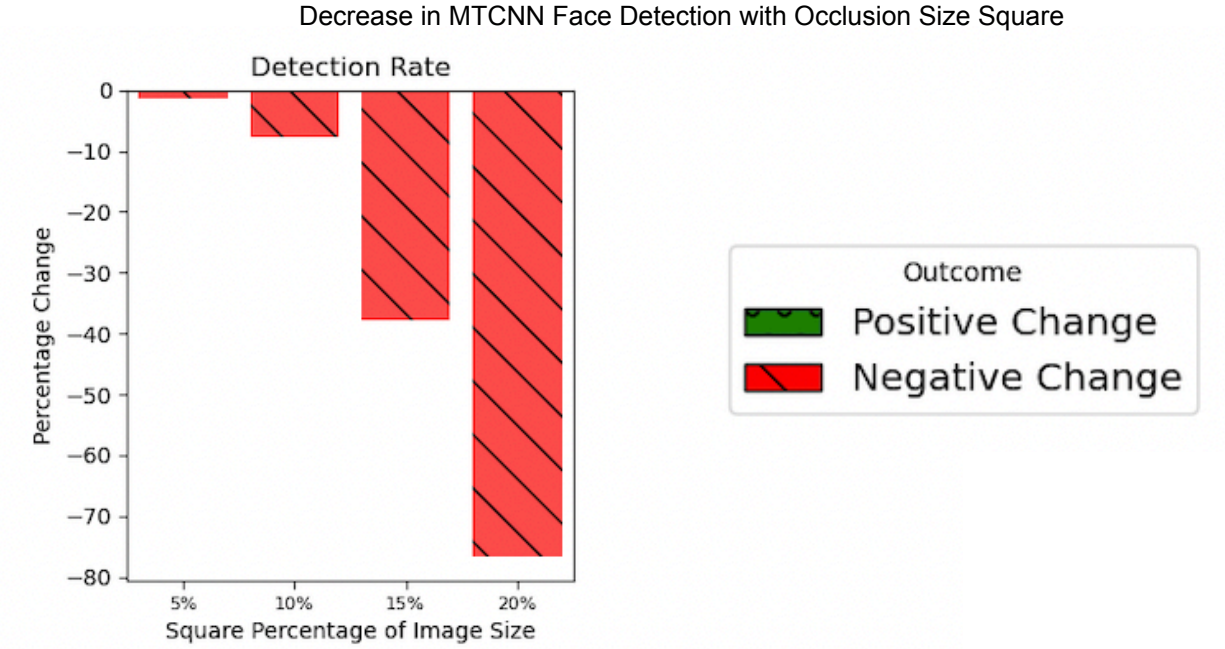


**Figure 16:** Percentage Change face detection with increase in Square size for Occlusion testing

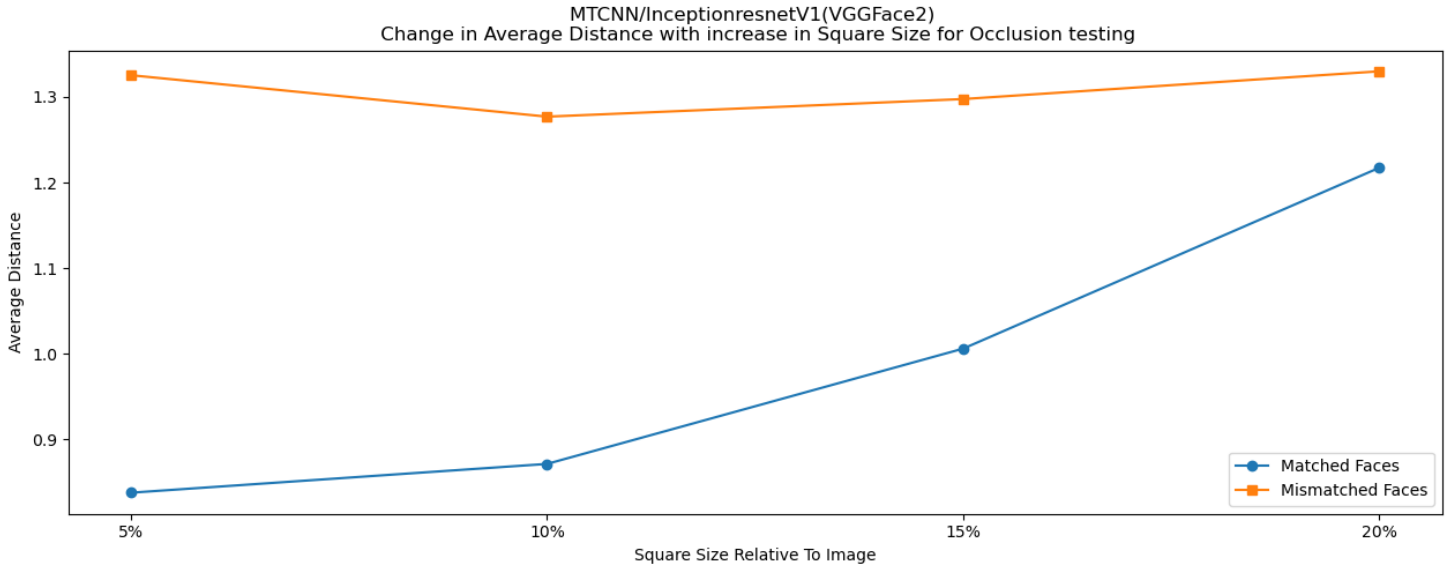


Figure 17: Change in distance values for the InceptionResNetV1(VGGFace2) as square size increases

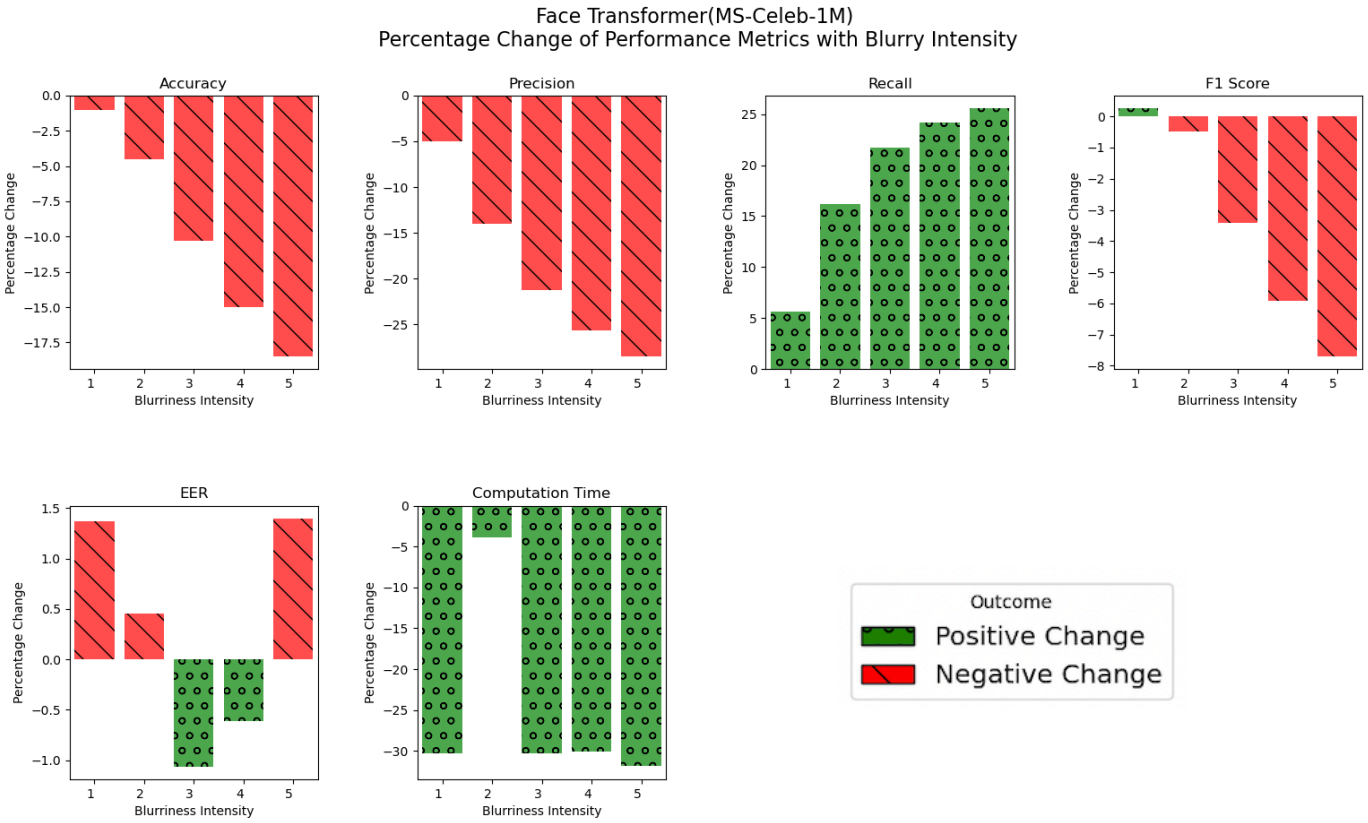


Figure 18: Percentage change of the Face Transformer metrics with increase in Blurry Intensity

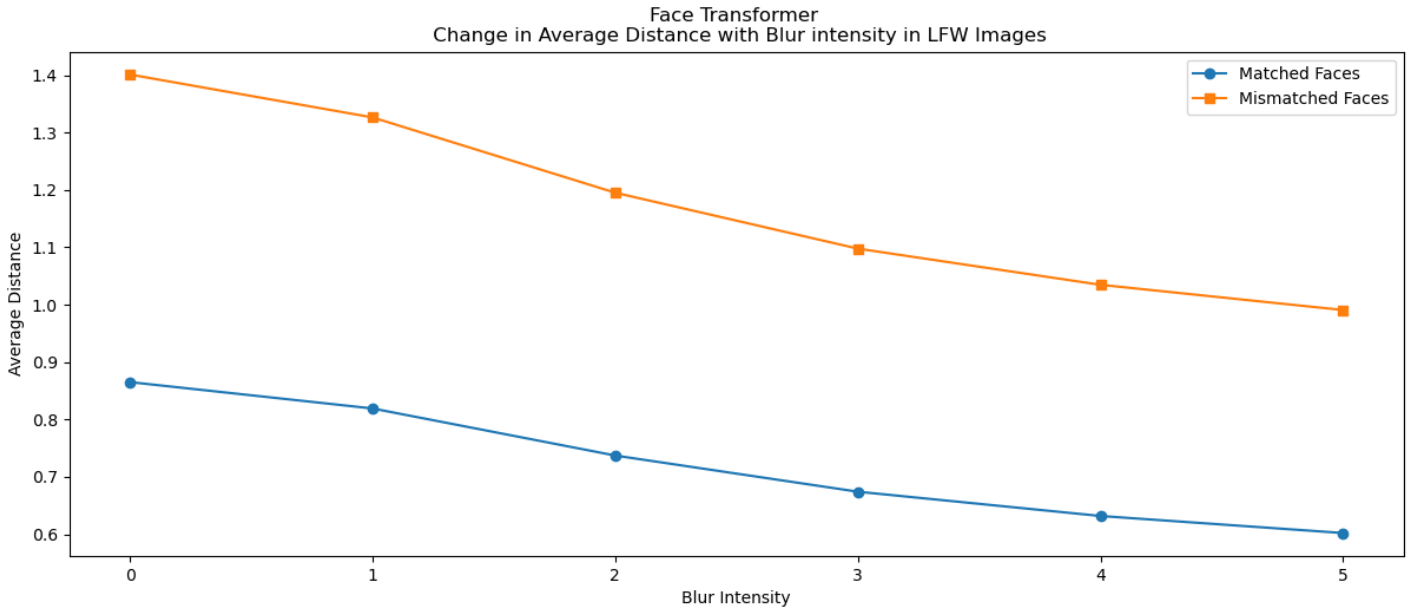


Figure 19: Change in distance values(Face Transformer) as square size for occlusion increases

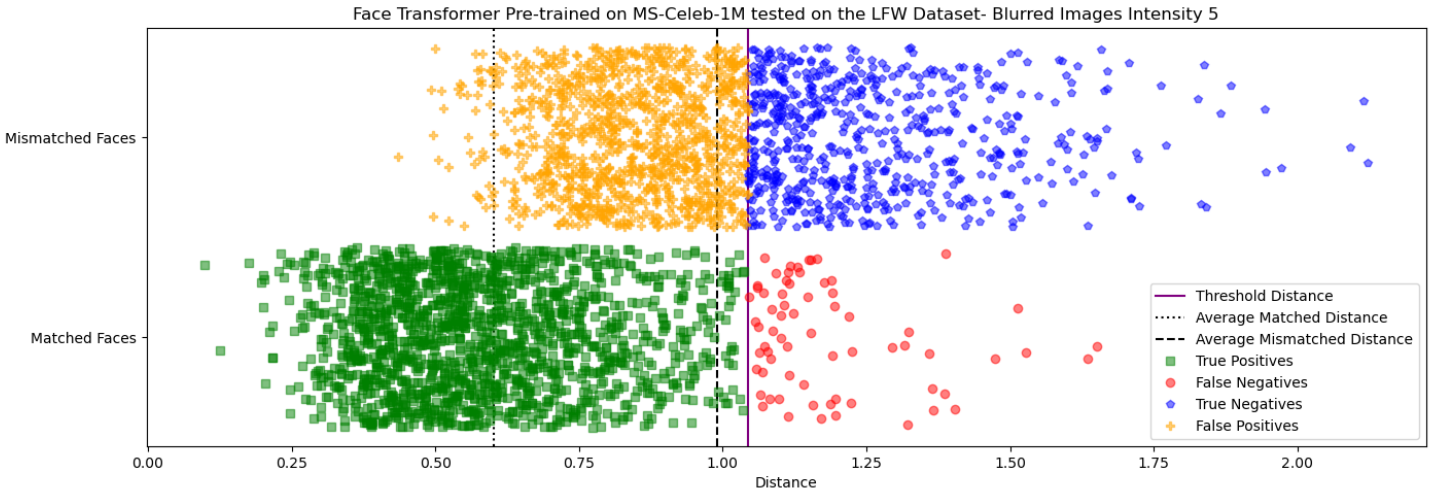


Figure 20: Plot illustrating TP, TN, FP and FN for the FaceTransformer - Blur Intensity 5

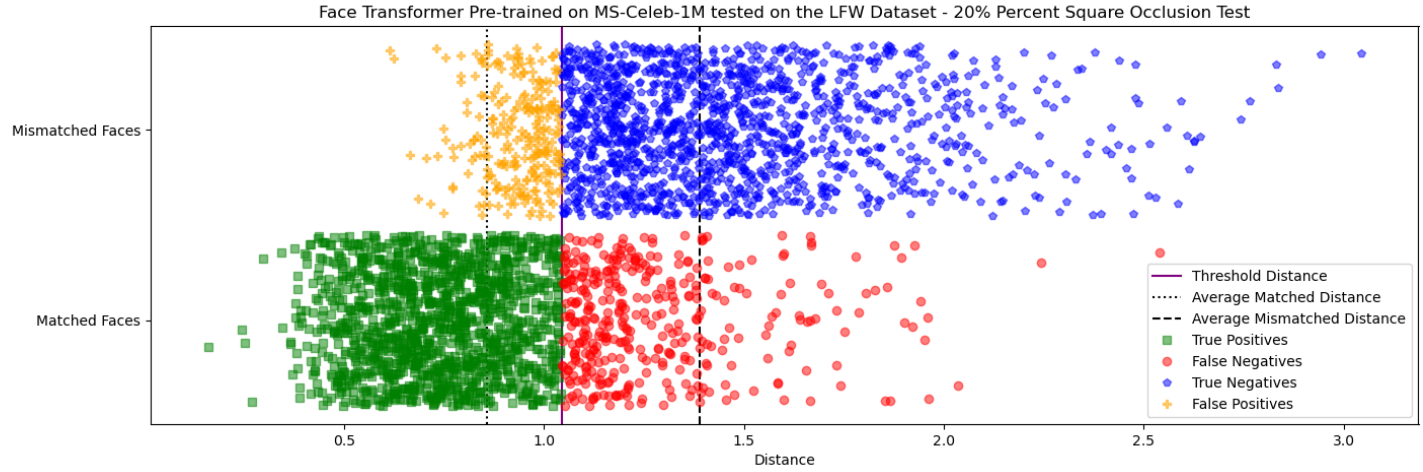


Figure 21: Plot illustrating TP, TN, FP and FN for the FaceTransformer - 20% Square Size for Occlusion

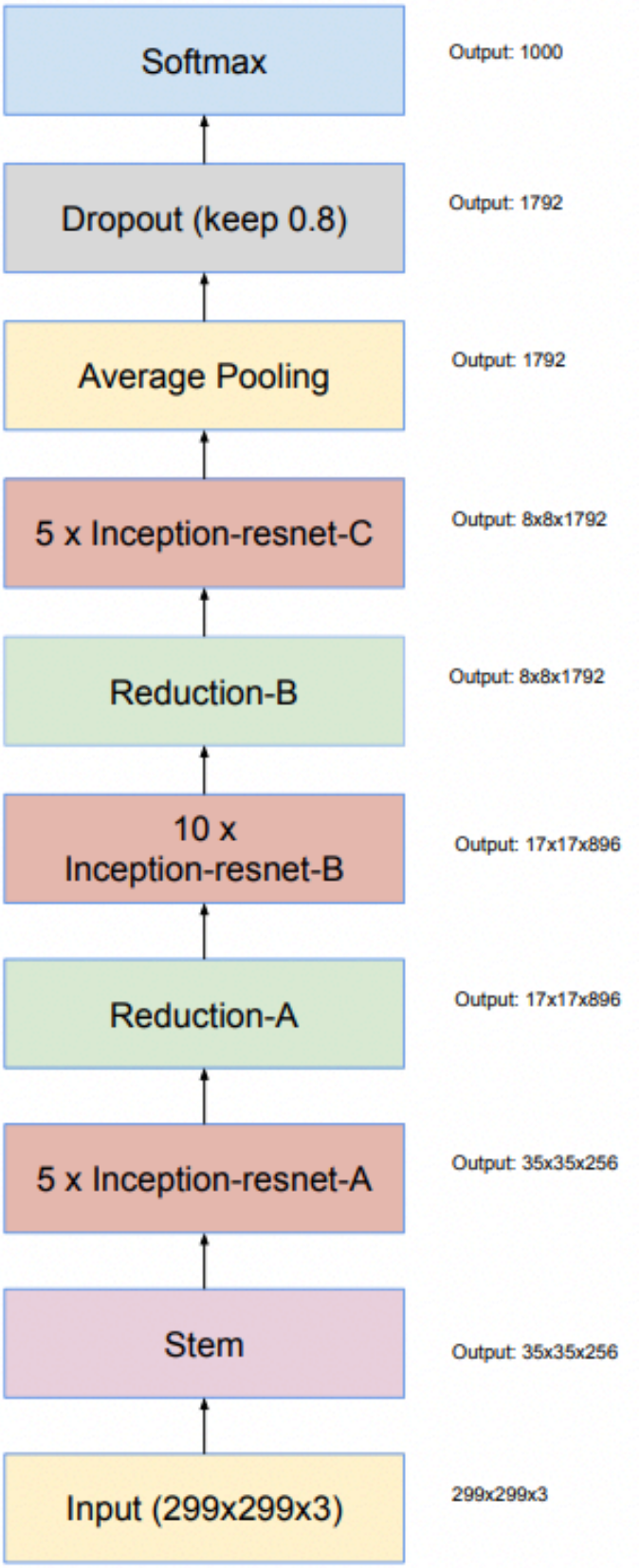


Figure 22: Schematic view for Inception-ResNet-v1. Figure taken from Szegedy et al. (2016)

**Department of Electronic, Electrical and Systems Engineering**

**UNIVERSITY OF BIRMINGHAM**

**GENERAL ETHICAL QUESTIONNAIRE FOR ALL STUDENTS**

| | |
|---|---|
| Name of student | Bauan Rashid |
| Email address of student | bxm857@student.bham.ac.uk |
| Name of supervisor | Dr Neil Cooke |

| | |
|---|---|
| Title of Research Project: | A Comparative Analysis of state-of-the-art Pre-Trained Face Recognition Models and the new Face Transformer on the LFW Dataset |

| | |
|---|---|
| **Will the research project involve humans as participants of the research** (with or without their knowledge or consent at the time)? This will include any survey, | No |
| **Are the results of the research project likely to expose any person to physical or psychological harm?** (Note, before starting the project you will need to complete a risk assessment in all cases) | No |
| **Will you have access to personal information that allows you to identify individuals, or to corporate or company confidential information** (that is not covered by confidentiality terms within an agreement or by a separate confidentiality agreement)? | No |
| **Does the research project present a significant risk to the environment or society?** | No |
| **Are there any ethical issues raised by this research project that in the opinion of your supervisor require further ethical review?** | No |

You have answered NO to all of the above questions. Further ethical review is not necessary.
You should have this form available at the bench inspections and include it in your final report.