Demographic Feature Recognition From Occluded Images Using Convolution Neural Networks

Neeraja Sanjay[a], Shreya Dhareshwar[a]

[a] *University at Buffalo, Department of Computer Science and Engineering, Buffalo, New York,*

Abstract

Deep learning approaches have demonstrated remarkable performance in various tasks such as object detection and face recognition. However, the detection of faces and face demographics in occluded images, such as images of individuals wearing masks or sunglasses, typically experience a decline in performance. In this project, we propose the development of three custom **CNN** architectures, based on **AlexNet**, **ResNet** and **VGG16** architectures, tailored for **multi-task deep learning** models, to address the challenge of accurately predicting gender, age, and ethnicity from a facial image even when a person's face is partially occluded. Our approach leverages convolutional neural network (CNN) based architectures to build deep learning models. Following the evaluation of the models, we further conduct experiments to assess their effectiveness on non-superimposed (real-time) **occluded** facial images and non-occluded facial images. This allows us to evaluate the models' generalization and robustness in real-world scenarios. The results demonstrate that our models exhibit improved performance compared to existing models when applied to the same case. These findings have significant implications in surveillance, marketing, and for other real-world applications where face detection and recognition are critical, particularly in environments where facial occlusion is common.

Keywords: CNN, AlexNet, ResNet, VGG16, multi-task deep learning, occluded

1. Introduction

The detection of faces and extraction of face demographics have gained significant attention due to their wide-ranging applications in fields such as computer vision, biometrics, artificial intelligence, and social sciences. Accurately analyzing attributes such as gender, age, and ethnicity from facial images has numerous practical implications, including surveillance, personalized user experiences, targeted advertising, forensic investigations, and demographic studies. However, the performance of face and face demographic detection models is known to decline in occluded images, where individuals wear masks, scarves, or sunglasses. Occlusions pose a considerable challenge for existing face detection models, as the presence of masks or other obstructions can hinder the accurate identification and analysis of facial features. The rise of face covering in various scenarios, such as during pandemics or in surveillance footage, has amplified the need for robust and reliable methods to detect and analyze faces, regardless of occlusions. The traditional approach to predicting face attributes involved constructing low-level descriptors through landmark detection and building domain classifiers based on these descriptors. This method relied on handcrafted features and separate classifiers for each attribute, which limited scalability, generalization, and adaptability to occluded or challenging facial images. Moreover, the existing models have not emphasized retraining. These models often utilized datasets consisting of black and white images and focused on detecting facial attributes from images with only one type of occlusion, such as a single kind of mask, which could skew the results when the model encounters an occluded image different from the pattern present in the training data.

The objective of this research paper is to address the limitations of existing models by developing improved deep-learning approaches for face demographics analysis in the presence of occlusions. Specifically, we propose leveraging convolutional neural network (CNN) based architectures that employ a multi-task learning approach to predict demographic features from occluded facial images. We utilize a repeated training strategy where we fine-tune the models and update their weights, thereby optimizing their performance.

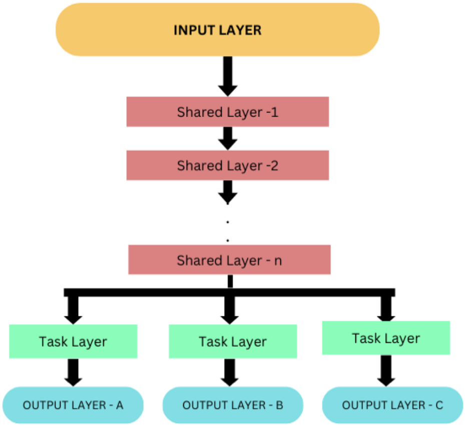


Figure 1: Multi-task learning Model Architecture

Throughout this paper, we will present our methodology, experimental setup, results and provide explainability into the model's decision-making process by employing visualizations of class activation maps to understand which regions of the face contribute most significantly to the predicted face demographics. We showcase the effectiveness and significance of our proposed models in addressing the challenges of face demographic detection in occluded images.

2. Related Work

2.1. *Prerana Mukherjee, Vinay Kaushik, Ronak Gupta, Ritika Jha, Daneshwari Kankanwadi, and Brejesh Lall: Attribute Prediction in Masked facial images with deep multitask learning: 9th International Conference on Pattern Recognition and Machine Intelligence (PReMI 2021) [1]*

The paper proposes a **multitask learning** approach where a deep neural network is trained to simultaneously predict multiple facial features, such as **age**, **gender**, and **ethnicity**, in the presence of masks. The authors investigate the effectiveness of leveraging shared representations among these tasks to improve the accuracy and robustness of feature prediction.

We have adopted the multitask learning approach presented in the paper to train a single network that predicts age, gender, and ethnicity simultaneously, even when faced with occlusions. We tried a variation of the resNet model used in the paper on resized **colored** occluded facial images of **50x50** resolution and included a **retraining** phase with which we were able to increase the accuracy of gender prediction from **0.8913** to **0.9183**, ethnicity from **0.7927** to **0.8203** and decreased avg. MAE of age from **12.15** to **5.5**

2.2. *Ruder, S.: An overview of multi-task learning in deep neural networks. arXiv preprint arXiv:1706.05098 (2017) [2]*

The paper primarily focuses on providing a comprehensive review and analysis of the multi-task learning (MTL) framework in deep neural networks. As such, the paper does not present specific experimental results or utilize a specific dataset for evaluation.

Instead, the paper primarily synthesizes and discusses existing literature and research findings related to MTL in deep neural networks. It provides insights into the motivations, benefits, challenges, and various approaches in the field of MTL. The author references and analyzes a wide range of studies from the MTL domain to provide a comprehensive overview of the topic.

We have incorporated concepts borrowed from this paper by using multi-task learning (**MTL**) for training a single network to predict age, gender, and ethnicity simultaneously, leveraging shared representations.

2.3. *G. Wu, J. Tao and X. Xu, "Occluded Face Recognition Based on the Deep Learning," 2019 Chinese Control and Decision Conference (CCDC), Nanchang, China, 2019, pp. 793-797, doi: 10.1109/CCDC.2019.8832330 [3]*

In the paper, the authors address the challenging task of recognizing occluded faces using deep learning techniques. They propose an approach that leverages the power of deep neural networks to extract discriminative features from occluded face images, thereby improving the accuracy of face recognition in the presence of occlusions.

We have adopted the deep learning-based approach proposed in the paper in our models and enhanced the performance by accurately predicting age, gender, and ethnicity despite the presence of occlusions in the facial images.

2.4. *Vikas Sheoran, Shreyansh Joshi, & Tanisha R. Bhayani (2021). Age and Gender Prediction using Deep CNNs and Transfer Learning. CoRR, abs/2110.12633.[4]*

In the paper, the authors focus on age and gender prediction tasks using deep convolutional neural networks (CNNs) and transfer learning techniques. They explore the effectiveness of pre-trained CNN models and investigate the impact of different architecture choices and transfer learning strategies for accurate age and gender prediction.

Specifically, the paper investigates different configurations of convolutional layers, pooling layers, and fully connected layers in the CNN architecture. The authors experiment with varying depths, filter sizes, and strides of the convolutional layers to capture different levels of visual features. They also explore the impact of different pooling operations, such as max pooling or average pooling, to downsample the feature maps and reduce spatial dimensions.

By adapting their architecture choices, we aim to improve the robustness of age and gender detection. Unlike the paper which uses **transfer learning** on non-occluded black and white images, we have trained the models from scratch on occluded colored images.

2.5. *Abrar H. Abdulnabi, Gang Wang, Jiwen Lu, & Kui Jia (2015). Multi-Task CNN Model for Attribute Prediction. IEEE Transactions on Multimedia, 17(11), 1949–1959.[5]*

The paper presents a **multi-task** convolutional neural network (CNN) model designed for attribute prediction. The authors recognize that attributes such as age, gender, and ethnicity can be simultaneously predicted from facial images. Therefore, they propose a unified framework that allows joint prediction of multiple attributes using a single CNN architecture.

The key contribution of the paper lies in the design of the multi-task CNN model. The authors leverage a **shared convolutional** feature extraction network that captures discriminative facial features across different attributes. This shared network is followed by individual task-specific branches that predict each attribute independently. The model is trained using a multi-task loss function that considers the correlation between different attributes, encouraging shared feature representations while maintaining attribute-specific prediction capabilities.

We have utilized the concept of using multi-task CNN model for attribute prediction inspired by the aforementioned paper. By adopting a similar framework, we are able to simultaneously predict gender, age, and ethnicity from facial images using a unified architecture. This allows for efficient and cohesive attribute prediction, leveraging shared convolutional features while preserving attribute-specific prediction capabilities.

3. Initial Approach

3.1. Algorithm 1: AlexNet variation

To begin with, we developed a custom CNN for multi-task learning with a structure like that of the AlexNet architecture. However, during our evaluation, we observed that this architecture did not yield satisfactory results, particularly in terms of ethnicity and age prediction. The model incorporated several convolutional layers, batch normalization, and LeakyReLU activation functions, allowing it to capture complex spatial features from partially occluded facial images. MaxPooling layers were employed to down-sample the feature maps and reduce spatial dimensions. The architecture also included separate branches for gender, ethnicity, and age prediction. While the model demonstrated a reasonable performance for gender, it exhibited limitations in accurately predicting ethnicity even after retraining thrice with different batch sizes and epoch values. Despite the inclusion of dense layers with ReLU activation and an appropriate loss function (categorical cross-entropy), the model struggled to capture the intricate patterns and nuances necessary for robust ethnicity and classification.

The results that were produced were unsatisfactory, so we determined that further refinement of the model was necessary. This led us to propose an enhanced architecture that specifically addressed the challenges associated with ethnicity and age classification.

3.2. Algorithm 2: ResNet variation

As an alternate approach, we developed a model which is based on the ResNet architecture presented in the state-of-the-art multi-task learning model[1]. The Novel elements of our version is that our model uses resized colored images of 50x50 resolution and incorporates retraining techniques. This enhanced model architecture resulted in improved performance compared to the AlexNet variation model.

Like the AlexNet variation model, the model architecture consists of several convolutional layers followed by batch normalization and leaky ReLU activations. Max pooling and average pooling layers are incorporated to extract essential features. The architecture employs three parallel branches, each consisting of convolutional and fully connected layers, with dropout regularization applied. The model outputs include gender (binary), ethnicity (categorical), and age (continuous) predictions. The model is compiled with the Adam optimizer and utilizes appropriate loss functions and metrics for each output. The proposed model has a total of 13,228,647 parameters, with 13,226,599 trainable parameters and 2,048 non-trainable parameters.

In the first training phase, the model was trained with a batch size of 8 for 39 epochs. In the second training phase (retraining), the model was retrained with a batch size of 12 for 4 epochs. The retraining process used the same configuration as the initial training, but with a smaller batch size and a shorter training duration.

This retraining aimed to further enhance the model's performance. This model demonstrated superior performance compared to the state-of-the-art model for all three attributes, gender, ethnicity, and age. The achieved accuracy rates were higher, indicating the effectiveness of the model architecture and the utilization of retraining techniques. While the achieved accuracy for the ethnicity attribute was commendable, there is still room for improvement. To further enhance the model's precision in predicting ethnicity, we propose another methodology.

4. Proposed Methodology

4.1. Dataset

To develop robust models for face demographic analysis, it is essential to have a diverse dataset that encompasses various age groups, genders, and ethnicities. This diversity ensures that the models can effectively capture the nuances and characteristics associated with different demographic attributes. Additionally, to address the challenges posed by occlusions, the dataset should include images with occlusions. To use face masks as a primary occlusion, the masked images should have a variety of masks that contain different textures, patterns, and colors to simulate real-world scenarios. By incorporating different types of masks, the dataset enables the models to learn and generalize to a wide range of occlusion variations commonly encountered in practice.

To address the first challenge, the **UTK Face Dataset** is used. The UTK Face Dataset consists of over 20,000 images of individuals from different ethnicities, including White, Black, Asian, Indian, and others. Each image is annotated with the person's age, gender, and ethnicity, making it a valuable resource for studying face demographics and related tasks. One notable aspect of the UTK Face Dataset is its focus on age estimation. The dataset provides a comprehensive age range, spanning from 0 to 116 years.

To incorporate masked images into our dataset, we employ a tool called MaskTheFace [6]. This tool generates masks with diverse texture and color variations, including popular mask types such as N95, cloth, and fiber masks. Moreover, it covers a wide range of face angles, ensuring that the generated masks are suitable for various facial orientations.

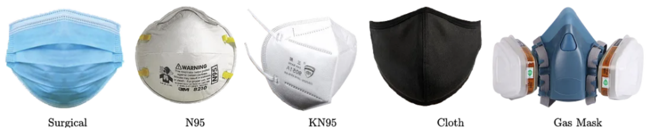


Figure 2: Different mask types

MaskTheFace is a computer vision-based script designed for masking faces in images. It utilizes a face landmarks detector based on the dlib library to accurately locate key facial features and determine the tilt of the face. The script includes a library of mask templates that are selected based on the face tilt. These templates are then transformed to fit the specific face by adjusting them according to the detected key facial features.

After superimposing the masks on the dataset, the resulting images are processed by flattening them to a standardized size.

This step is crucial to ensure consistency in the input data for subsequent analysis. The flattened images, which have undergone this resizing procedure, are then paired with their corresponding age, gender, and ethnicity labels.

To organize and manage this data effectively, a data frame structure is employed. The data frame consists of separate columns dedicated to age, gender, ethnicity, and flattened image data, which has been resized to a resolution of 50x50 pixels.

The dataset is divided such that 60% of the data is allocated to the training set, 20% to the testing set, and the remaining 20% to the validation set. This distribution allows for a substantial amount of data for model training, while also providing sufficient samples for rigorous testing and unbiased evaluation.

4.2. Algorithm

We propose a model architecture that is a variation of the VGG16 model, specifically designed for multi-task learning. It aims to predict gender, ethnicity, and age simultaneously from colored input images. prediction. Each component follows a specific set of layers and operations to learn the corresponding attribute. This multi-task learning architecture employs hierarchical parameter sharing, where a common encoder module is shared among all tasks. This approach enables the model to learn shared representations while also focusing on task-specific features. By jointly learning these 3 attributes, the model can leverage the interdependencies among them, leading to improved performance across all tasks.

The gender prediction, ethnicity prediction, and age prediction components of the model share a common encoder architecture, which includes convolutional layers with batch normalization and leaky ReLU activation. This allows the model to extract relevant features from the input image. Additionally, all three components employ additional convolutional layers and dense layers with ReLU activation to capture more complex patterns and learn discriminative representations. The model was initially trained for 40 epochs using a batch size of 32.

It was then retrained twice, with smaller batch sizes and fewer epochs. In the first retraining session, the model was trained for 5 epochs with a batch size of 17. In the second retraining session, the model was trained for 3 epochs with a batch size of 32. These retraining sessions aimed to fine-tune the model's performance and improve its predictions for gender, ethnicity, and age from input images.

In terms of differences, the gender prediction component utilizes max pooling after feature extraction, whereas the ethnicity prediction component uses average pooling. This leads to different down-sampling techniques and can affect the spatial dimensions of the extracted features.

Furthermore, the output layers differ across the components. The gender prediction component has a dense layer with a sigmoid activation function, producing a binary output (0 or 1) representing male or female. The ethnicity prediction component, on the other hand, applies a softmax activation function in its final dense layer, yielding a categorical output with 5 classes representing different ethnicities (0 – white, 1 – black, 2 - Asian, 3 – Indian, 4 – Other).

Lastly, the age prediction component predicts the age as a continuous value using a final dense layer without any specific activation function. The specific differences in pooling, output layers, and target variables allow the model to focus on task-specific characteristics.

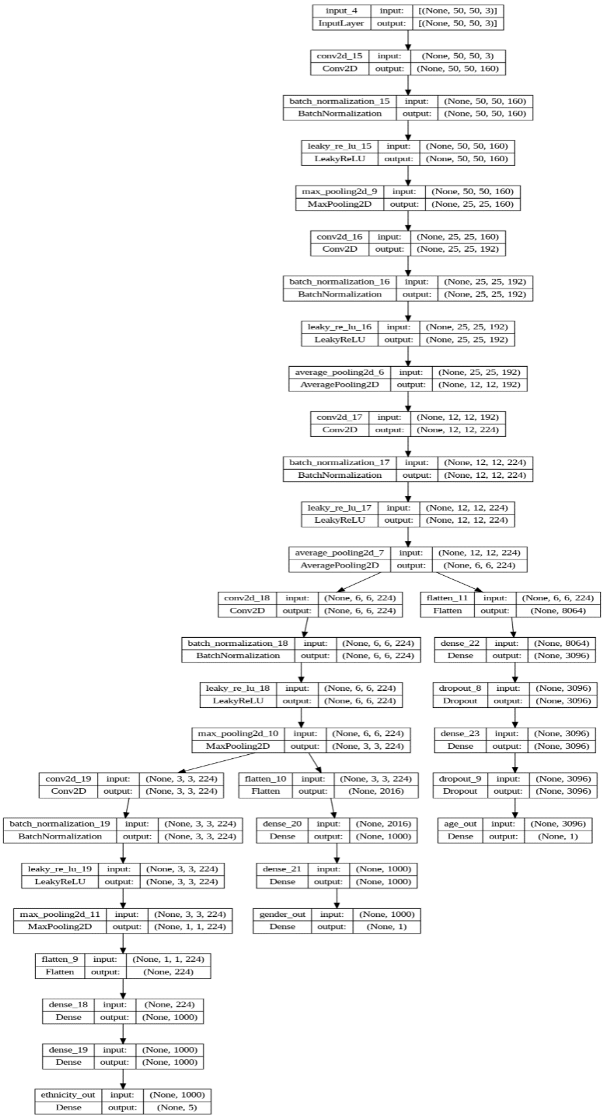

input_4 input: [(None, 50, 50, 3)]
InputLayer output: [(None, 50, 50, 3)]
conv2d_15 input: (None, 50, 50, 3)
Conv2D output: (None, 50, 50, 160)
batch_normalization_15 input: (None, 50, 50, 160)
BatchNormalization output: (None, 50, 50, 160)
leaky_re_lu_15 input: (None, 50, 50, 160)
LeakyReLU output: (None, 50, 50, 160)
max_pooling2d_9 input: (None, 50, 50, 160)
MaxPooling2D output: (None, 25, 25, 160)
conv2d_16 input: (None, 25, 25, 160)
Conv2D output: (None, 25, 25, 192)
batch_normalization_16 input: (None, 25, 25, 192)
BatchNormalization output: (None, 25, 25, 192)
leaky_re_lu_16 input: (None, 25, 25, 192)
LeakyReLU output: (None, 25, 25, 192)
average_pooling2d_6 input: (None, 25, 25, 192)
AveragePooling2D output: (None, 12, 12, 192)
conv2d_17 input: (None, 12, 12, 192)
Conv2D output: (None, 12, 12, 224)
batch_normalization_17 input: (None, 12, 12, 224)
BatchNormalization output: (None, 12, 12, 224)
leaky_re_lu_17 input: (None, 12, 12, 224)
LeakyReLU output: (None, 12, 12, 224)
average_pooling2d_7 input: (None, 12, 12, 224)
AveragePooling2D output: (None, 6, 6, 224)
conv2d_18 input: (None, 6, 6, 224)
Conv2D output: (None, 6, 6, 224)
flatten_11 input: (None, 6, 6, 224)
Flatten output: (None, 8064)
batch_normalization_18 input: (None, 6, 6, 224)
BatchNormalization output: (None, 6, 6, 224)
dense_22 input: (None, 8064)
Dense output: (None, 3096)
leaky_re_lu_18 input: (None, 6, 6, 224)
LeakyReLU output: (None, 6, 6, 224)
dropout_8 input: (None, 3096)
Dropout output: (None, 3096)
max_pooling2d_10 input: (None, 6, 6, 224)
MaxPooling2D output: (None, 3, 3, 224)
dense_23 input: (None, 3096)
Dense output: (None, 3096)
conv2d_19 input: (None, 3, 3, 224)
Conv2D output: (None, 3, 3, 224)
flatten_10 input: (None, 3, 3, 224)
Flatten output: (None, 2016)
dropout_9 input: (None, 3096)
Dropout output: (None, 3096)
batch_normalization_19 input: (None, 3, 3, 224)
BatchNormalization output: (None, 3, 3, 224)
dense_20 input: (None, 2016)
Dense output: (None, 1000)
age_out input: (None, 3096)
Dense output: (None, 1)
leaky_re_lu_19 input: (None, 3, 3, 224)
LeakyReLU output: (None, 3, 3, 224)
dense_21 input: (None, 1000)
Dense output: (None, 1000)
max_pooling2d_11 input: (None, 3, 3, 224)
MaxPooling2D output: (None, 1, 1, 224)
gender_out input: (None, 1000)
Dense output: (None, 1)
flatten_9 input: (None, 1, 1, 224)
Flatten output: (None, 224)
dense_18 input: (None, 224)
Dense output: (None, 1000)
dense_19 input: (None, 1000)
Dense output: (None, 1000)
ethnicity_out input: (None, 1000)
Dense output: (None, 5)

Figure 3: VGG16 Model

To train the model, appropriate loss functions are applied for each task: binary cross-entropy for gender prediction, categorical cross-entropy for ethnicity prediction, and mean squared error for age prediction. The model's performance is evaluated using metrics such as accuracy for gender and ethnicity predictions and mean absolute error (MAE) for age estimation. The total number of parameters in the proposed model architecture is 40,386,815. Out of the total parameters, 40,384,767 are trainable parameters. These are the parameters that are updated through backpropagation during training, allowing the model to learn and adjust its internal representations based on the given data. On the other hand, there are 2,048 non-trainable parameters. These parameters are typically associated with fixed components within the model, such as batch normalization layers or other non-trainable operations. Non-trainable parameters do not get updated during the training process and are usually set during the model's initialization. Class Activation Maps (CAM) are utilized to provide interpretability to the prediction models. CAMs highlight the regions in the input image that contribute the most to the final prediction, allowing us to understand which areas of the image are influential for each task.

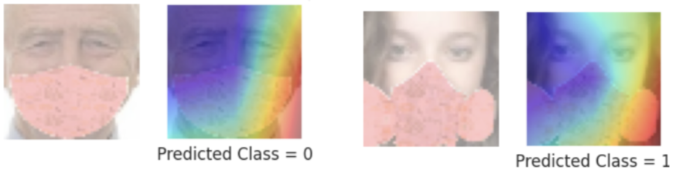


Figure 4: Gender Class Activation Map

Red regions denote where maximum activation is attained. From Fig(4) we can infer that in the case of gender prediction, the CAM can provide insights into the features that distinguish between males and females. Based on the provided information, it appears that the ear region plays a significant role in gender classification. If the ear region is visible in the input image, the model may focus on that area to make its prediction. For example, if the ear region is prominently detected, the model may classify it as male. On the other hand, if there is hair covering the side of the forehead(bangs) and ear, the model might infer it as a characteristic more commonly associated with women.

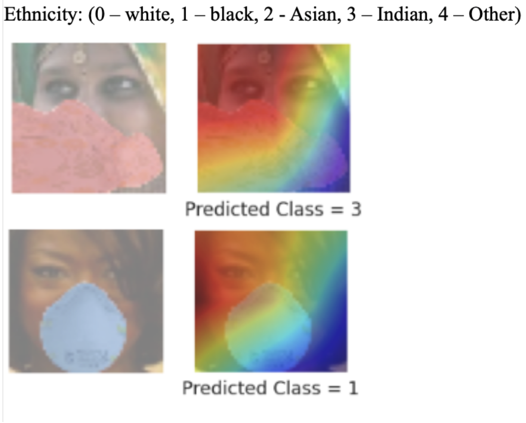

Ethnicity: (0 – white, 1 – black, 2 - Asian, 3 – Indian, 4 – Other)
Predicted Class = 3
Predicted Class = 1

Figure 5: Ethnicity Class Activation Map

In the context of ethnicity-based prediction, it was significant that the forehead region plays a crucial role in determining ethnicity with reference to Fig(5). The forehead region has been found to carry significant discriminatory information for ethnicity prediction. It exhibits distinctive characteristics that can help differentiate between different ethnic groups. Forehead features may include variations in skin tone, texture, or the presence of specific facial structures. For instance, if the CAM highlights the forehead region and detects the presence of a "bindhi" (a decorative mark worn by some Indian women on the forehead), the model may associate it as a distinguishing feature for classifying the individual as Indian.

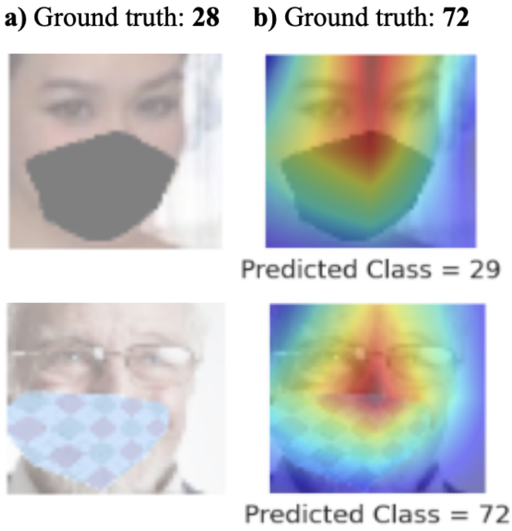


Figure 6: Age Class Activation Map

For age prediction, Fig(6) shows that the model highlights the bridge of the nose and the middle of the forehead. The bridge of the nose and the middle of the forehead are areas of the face that commonly exhibit visible signs of aging, such as wrinkles, lines, and texture changes. These regions are prone to showing age-related characteristics due to factors like sun exposure, muscle movement, and collagen loss. If the CAM highlights the bridge of the nose and the middle of the forehead and detects the presence of many wrinkles and lines, the model may predict the individual of belonging to an older age group.

This model shows significant advancements compared to our previous models and has surpassed the **state-of-the-art** model in terms of performance for detecting age, gender, and ethnicity in occluded facial images. In addition to this, the proposed model architecture also demonstrates good accuracy in detecting these features in non-occluded images and real-time images involving individuals wearing non-superimposed masks or sunglasses. The high accuracy achieved by the model is a testament to its robustness and generalization capabilities. It demonstrates the model's ability to effectively learn and utilize facial features that are indicative of gender, ethnicity, and age, even in challenging and realistic conditions.

5. Results

5.1. Quantitative Analysis

In terms of gender accuracy, the VGG16 Variation model achieves the highest accuracy of 0.9509, followed by the ResNet Variation model with an accuracy of 0.9182. The AlexNet model has a slightly lower gender accuracy of 0.7943. For ethnicity accuracy, the VGG16 Variation model again performs the best with an accuracy of 0.8733, followed by the ResNet Variation model with an accuracy of 0.8203. The AlexNet model has the lowest ethnicity accuracy of 0.5044.

Model Architecture &	Type of Images	Resolution	Gender (Acc.)	Ethnicity (Acc.)	Age (MAE)
ResNet (SOTA)	**B&W**	**48x48**	**0.8913**	**0.7927**	**12.15**
AlexNet	Color	50x50	0.7943	0.5044	8.209
ResNet Variation	Color	50x50	0.9182	0.8203	5.537
VGG16 Variation (Optimal)	**Color**	**50x50**	**0.9509**	**0.8733**	**6.598**

Table 1: Performance of Different Models on Attribute Prediction

In terms of age MAE, the ResNet Variation model achieves the lowest MAE of 5.537, indicating better age estimation accuracy. The VGG16 Variation model has a slightly higher MAE of 6.598, while the AlexNet model has the highest MAE of 8.209.

Overall, the VGG16 Variation model showcased a balanced performance across all evaluated metrics, with competitive accuracy in gender, ethnicity, and age prediction tasks. Its consistent and reliable performance led us to conclude that the VGG16 Variation model is the most optimal choice for our task at hand.

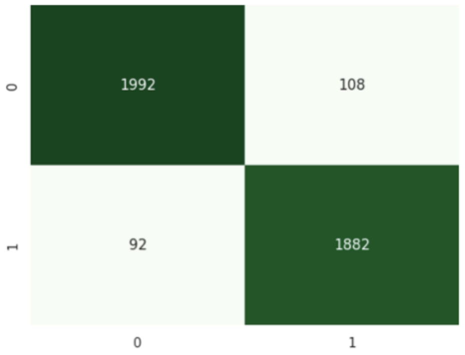


Figure 7: Confusion Matrix - Gender

Class	Precision	Recall	F1 score
Male	0.96	0.95	0.95
Female	0.95	0.95	0.95

Table 2: Gender Classification

These results indicate that the model has high accuracy in predicting both male and female genders, with balanced precision, recall, and F1-score values for both classes.

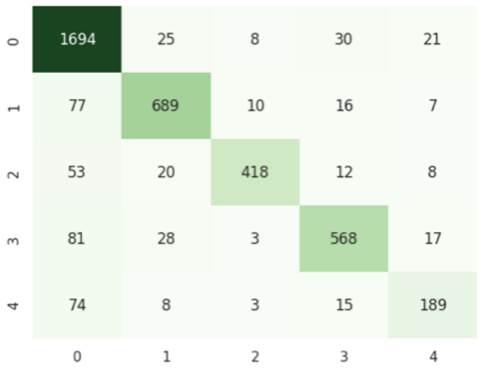


Figure 8: Confusion Matrix - Ethnicity

Class	Precision	Recall	F1 Score
White	0.86	0.95	0.90
Black	0.89	0.86	0.88
Asian	0.95	0.82	0.88
Indian	0.89	0.81	0.85
Others	0.78	0.65	0.71

Table 3: Ethnicity Classification

For the ethnicity classes (white, black, Asian, and Indian), the model demonstrates relatively higher precision values ranging from 0.86 to 0.95, indicating good accuracy in predicting samples from these ethnicities. The recall values range from 0.81 to 0.95, suggesting that the model is effective at identifying a large portion of the actual samples from these classes. The F1 scores for these ethnicities range from 0.85 to 0.90, indicating a balanced performance between precision and recall. Overall, the model shows promising results in predicting the ethnicity of individuals belonging to these specific classes. The scores for the ”others” class indicate that the model has difficulty correctly identifying samples from this category.

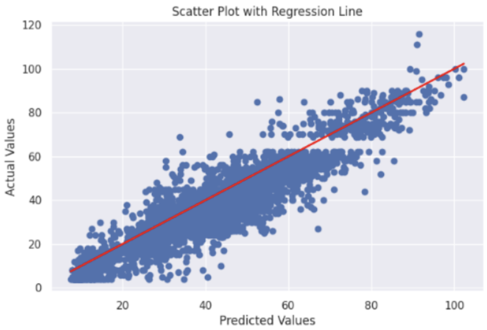


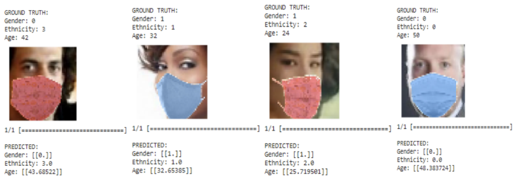
Figure 9: Scatter Plot - Age

Even though there is a slight deviation or scatter around the line, a strong overall correlation between the actual and predicted values suggests that the model captures the underlying patterns and trends in the age data. This indicates that the model can provide accurate age predictions, with some level of variability. The regression plot represents the relationship between the actual age values and the predicted age values by the model. In an ideal scenario, where the predictions perfectly align with the actual values, the regression plot would show a straight line with a slope of 1. This would indicate that the model's predictions are highly accurate and consistent with the true age values.

5.2. *Comparison with State-of-the-art*

Comparing our ResNet variation to the state-of-the-art ResNet, we observe improvements in all components with respect to Table 1. This implies that our ResNet model provides more precise age predictions, resulting in smaller deviations from the actual age values. Our VGG16 variation, on the other hand, exhibits even more impressive results compared to the state-of-the-art model and our ResNet model. The scores from Table 1 suggest that this VGG16 model provides the most optimal results. One important aspect to consider is the resolution of the images used to train the model. In our case, the VGG16 variation and the ResNet variation were trained on colored images with a resolution of 50x50 pixels, while the state-of-the-art ResNet was trained on black and white images with a resolution of 48x48 pixels. Even though we used the same architecture as the state-of-the-art model to develop our ResNet model, the superior performance of our ResNet and VGG16 variation models can be attributed to various factors, such as the utilization of colored images with a slightly higher resolution, the use of colored images, retraining of the models to adapt to the specific task, and potentially optimized model architectures. These factors enable our models to leverage richer visual information, learn task-specific features, and improve overall prediction accuracy.

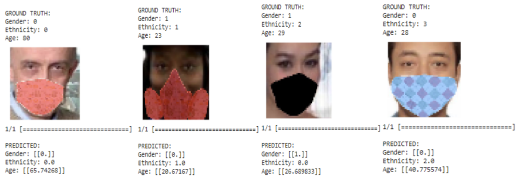
5.3. *Interpretation*



```
GROUND TRUTH:
Gender: 0
Ethnicity: 3
Age: 42
1/1 [==============================]
PREDICTED:
Gender: [[0.]]
Ethnicity: 3.0
Age: [[43.68522]]

GROUND TRUTH:
Gender: 1
Ethnicity: 1
Age: 32
1/1 [==============================]
PREDICTED:
Gender: [[1.]]
Ethnicity: 1.0
Age: [[32.65385]]

GROUND TRUTH:
Gender: 1
Ethnicity: 2
Age: 24
1/1 [==============================]
PREDICTED:
Gender: [[1.]]
Ethnicity: 2.0
Age: [[25.719501]]

GROUND TRUTH:
Gender: 0
Ethnicity: 0
Age: 50
1/1 [==============================]
PREDICTED:
Gender: [[0.]]
Ethnicity: 0.0
Age: [[48.383724]]
```

Figure 10: Success Cases - VGG16 variation

The images above showcase a subset of the test images and their correctly predicted results by the VGG16 variation model. Each label is paired with its corresponding ground truth label and predicted label by the model. Given that the Mean Absolute Error (MAE) of the age predictions by our VGG16 variation model is 6.598, we consider age predictions within a range of plus or minus 5 years (threshold) from the ground truth to be valid. This means that if the predicted age falls within this range, it is considered a reasonably accurate estimation.



```
GROUND TRUTH:
Gender: 0
Ethnicity: 0
Age: 80
1/1 [==============================]
PREDICTED:
Gender: [[0.]]
Ethnicity: 0.0
Age: [[65.74268]]

GROUND TRUTH:
Gender: 1
Ethnicity: 1
Age: 23
1/1 [==============================]
PREDICTED:
Gender: [[0.]]
Ethnicity: 1.0
Age: [[20.67167]]

GROUND TRUTH:
Gender: 1
Ethnicity: 2
Age: 29
1/1 [==============================]
PREDICTED:
Gender: [[1.]]
Ethnicity: 0.0
Age: [[26.689833]]

GROUND TRUTH:
Gender: 0
Ethnicity: 3
Age: 28
1/1 [==============================]
PREDICTED:
Gender: [[0.]]
Ethnicity: 2.0
Age: [[40.775574]]
```

Figure 11: Failure Cases - VGG16 variation

The ethnicity errors in the prediction of Asian facial images, where they are misclassified as White ethnicity, and the misclassification between Indian and Asian ethnicities can be observed. These errors indicate a challenge in accurately predicting the ethnicity attribute for certain individuals. The model tends to assign the White ethnicity label to Asian faces, possibly due to shared facial features and visual similarities between these groups. Additionally, there is confusion between Indian and Asian ethnicities, suggesting a difficulty in distinguishing subtle differences in facial characteristics.

We have identified age prediction errors that occur in facial images of individuals around the age of 75, particularly when they are bald. Notably, these errors manifest as the model predicting a lesser age than the actual age of the individual. One possible explanation for these errors is the presence of age-related visual cues that may not be adequately captured by the model. Facial aging can be influenced by various factors, including wrinkles, fine lines, and sagging skin, which may be less prominent in bald individuals or individuals with a smoother complexion as only the bridge of the nose and the middle of the forehead is used by the model to detect the age. Consequently, the model struggles to accurately estimate the age of these individuals, leading to the observed errors.

The model achieves low gender prediction errors, but a few errors still occur. Since the number of wrong gender classifications is very low, identifying the cause of these few errors is difficult. One possible explanation for these errors could be a skew in the data distribution.

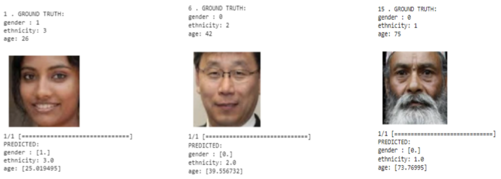
5.4. *Experiments*

5.4.1. *Testing on Non-Masked Facial Images*

Model Architecture	Method	Image Type	Result	Gender (Accuracy)	Ethnicity (Accuracy)	Age (MAE)
ResNet (SOTA)	**Existing**	**B&W**	**48x48**	**0.89**	**0.76**	**11.57**
ResNet Variation	Ours	Color.	50x50	0.94	0.70	5.63
VGG16 Variation (Optimal)	**Ours**	**Color.**	**50x50**	**0.95**	**0.89**	**5.68**

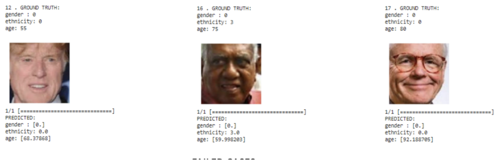
Table 4: Model Performance

From Table 4, we can infer that our VGG16 variation model showcased superior performance in non-masked facial images as well compared to the **state-of-the-art** model. We observed that the accuracy of gender, accuracy of ethnicity prediction, and the mean absolute error of age estimation was higher in non-masked facial images compared to masked images. This can be attributed to several factors. By removing the masks in non-masked images, the model has access to more complete and unobstructed facial information, allowing for better prediction accuracy.



```
1 . GROUND TRUTH:
gender : 1
ethnicity: 3
age: 26
1/1 [==============================]
PREDICTED:
gender : [1.]
ethnicity: 3.0
age: [25.019495]

6 . GROUND TRUTH:
gender : 0
ethnicity: 2
age: 42
1/1 [==============================]
PREDICTED:
gender : [0.]
ethnicity: 2.0
age: [39.556732]

15 . GROUND TRUTH:
gender : 0
ethnicity: 1
age: 75
1/1 [==============================]
PREDICTED:
gender : [0.]
ethnicity: 1.0
age: [73.76995]
```

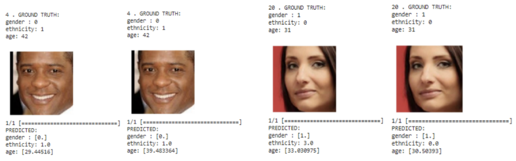
Figure 12: Success Cases - Non-Masked Facial images



```
12 . GROUND TRUTH:
gender : 0
ethnicity: 0
age: 55
1/1 [==============================]
PREDICTED:
gender : [0.]
ethnicity: 0.0
age: [68.37868]

16 . GROUND TRUTH:
gender : 0
ethnicity: 3
age: 75
1/1 [==============================]
PREDICTED:
gender : [0.]
ethnicity: 3.0
age: [59.998203]

17 . GROUND TRUTH:
gender : 0
ethnicity: 0
age: 80
1/1 [==============================]
PREDICTED:
gender : [0.]
ethnicity: 0.0
age: [92.188705]
```

Figure 13: Failure Cases - Non-Masked Facial images

In the evaluation of non-masked facial images, we observed some errors that were concentrated in age prediction, while the overall number of incorrect predictions was relatively low. One possible reason for this trend is the inherent complexity and variability in age-related features and appearances, for example, someone who is wearing professional makeup, could look younger than they are. These factors can contribute to the margin of error in age prediction, even when using sophisticated deep-learning models.



```
4 . GROUND TRUTH:
gender : 0
ethnicity: 1
age: 42
1/1 [==============================]
PREDICTED:
gender : [0.]
ethnicity: 1.0
age: [29.44516]
```

ResNet

```
4 . GROUND TRUTH:
gender : 0
ethnicity: 1
age: 42
1/1 [==============================]
PREDICTED:
gender : [0.]
ethnicity: 1.0
age: [39.483364]
```

VGG

```
20 . GROUND TRUTH:
gender : 1
ethnicity: 0
age: 31
1/1 [==============================]
PREDICTED:
gender : [1.]
ethnicity: 3.0
age: [33.030975]
```

ResNet

```
20 . GROUND TRUTH:
gender : 1
ethnicity: 0
age: 31
1/1 [==============================]
PREDICTED:
gender : [1.]
ethnicity: 0.0
age: [30.50393]
```

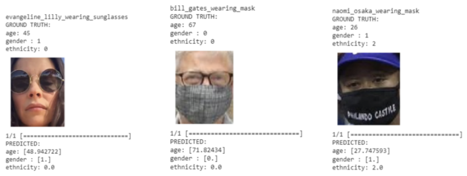
VGG

Figure 14: Comparison between ResNet variation and VGG16 variation - Non-Masked Facial images

While our ResNet variation demonstrates strong performance in gender prediction and overall ethnicity classification for non-masked facial images, it does exhibit some errors in age estimation and fine-grained ethnicity classification compared to our VGG16 variation. From Figure 14 we can infer that the VGG16 variation model predicts the age of the person in the image much closer to the ground truth age, unlike the Resnet variation. We can also notice that ResNet mispredicts the ethnicity. This can be attributed to the VGG16 model's ability to capture and learn more detailed and discriminative features specific to different ethnicities. The deeper architecture of VGG16 enables it to extract more intricate features related to facial structure, skin tone, and other ethnic-specific characteristics, leading to improved ethnicity classification accuracy. These results clearly demonstrate that our VGG16 variation model outperformed the state-of-the-art model in terms of gender prediction, ethnicity prediction, and age estimation in both masked and non-masked images.

5.4.2. *Testing on Real-time occluded facial images*

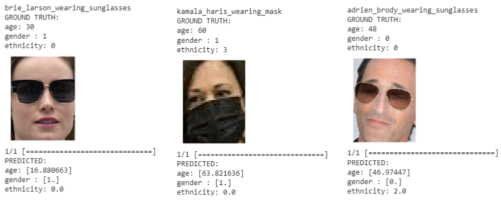
To assess the real-world applicability and robustness of our trained model, it is crucial to evaluate its performance on images that do not have superimposed occlusions. While the model has been trained on data with superimposed occlusions to simulate realistic scenarios, it is important to verify if it can maintain good accuracy and generalization when faced with non-superimposed occlusion images. To evaluate the model's performance on facial images that do not have superimposed occlusions, we used a subset of the "Real World Occluded Faces (ROF)" dataset. This dataset contains face image samples of celebrities with real-life upper-face and lower-face occlusions (i.e., face masks and sunglasses). All the images are from real-life scenarios and have large variations in pose and illumination.



```
evangeline_lilly_wearing_sunglasses
GROUND TRUTH:
age: 45
gender : 1
ethnicity: 0
1/1 [==============================]
PREDICTED:
age: [48.942722]
gender : [1.]
ethnicity: 0.0

bill_gates_wearing_mask
GROUND TRUTH:
age: 67
gender : 0
ethnicity: 0
1/1 [==============================]
PREDICTED:
age: [71.82434]
gender : [0.]
ethnicity: 0.0

naomi_osaka_wearing_mask
GROUND TRUTH:
age: 26
gender : 1
ethnicity: 2
1/1 [==============================]
PREDICTED:
age: [27.747593]
gender : [1.]
ethnicity: 2.0
```

Figure 15: Success Cases - Real-Time Occluded Facial images

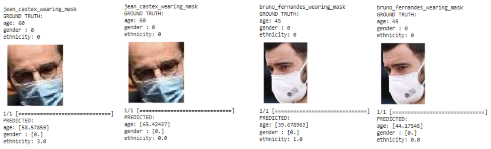
The VGG16 variation model demonstrated its efficacy even in this challenging scenario. It was able to correctly predict the facial attributes of images in the ROF dataset, including those with occlusions. This indicates that the model possesses robust features and can effectively capture and analyze facial characteristics, even when non-superimposed real occlusions are present. Out of the 50 images that were used from the ROF dataset, the model predicted all three attributes of 41 of them correctly, giving a cumulative accuracy of 82% for this task.



```
brie_larson_wearing_sunglasses
GROUND TRUTH:
age: 30
gender : 1
ethnicity: 0
1/1 [==============================]
PREDICTED:
age: [16.880663]
gender : [1.]
ethnicity: 0.0

kamala_haris_wearing_mask
GROUND TRUTH:
age: 60
gender : 1
ethnicity: 3
1/1 [==============================]
PREDICTED:
age: [63.821636]
gender : [1.]
ethnicity: 0.0

adrien_brody_wearing_sunglasses
GROUND TRUTH:
age: 48
gender : 0
ethnicity: 0
1/1 [==============================]
PREDICTED:
age: [46.97447]
gender : [0.]
ethnicity: 2.0
```

Figure 16: Failure Cases - Real Time Occluded Facial images

A notable number of error cases were observed specifically for sunglass occlusions. Sunglasses occluded the bridge of the nose. As this is a primary feature used by the model to detect the age, errors in age prediction occur. Errors in ethnicity prediction are caused primarily because the images are not of good clarity and are blurred.



```
jean_castex_wearing_mask
GROUND TRUTH:
age: 60
gender : 0
ethnicity: 0
1/1 [==============================]
PREDICTED:
age: [58.57859]
gender : [0.]
ethnicity: 3.0
```

RESNET

```
jean_castex_wearing_mask
GROUND TRUTH:
age: 60
gender : 0
ethnicity: 0
1/1 [==============================]
PREDICTED:
age: [65.42437]
gender : [0.]
ethnicity: 0.0
```

VGG

```
bruno_fernandes_wearing_mask
GROUND TRUTH:
age: 45
gender : 0
ethnicity: 0
1/1 [==============================]
PREDICTED:
age: [39.678963]
gender : [0.]
ethnicity: 1.0
```

RESNET

```
bruno_fernandes_wearing_mask
GROUND TRUTH:
age: 45
gender : 0
ethnicity: 0
1/1 [==============================]
PREDICTED:
age: [44.17545]
gender : [0.]
ethnicity: 0.0
```

VGG

Figure 17: Comparison between ResNet variation and VGG16 variation - Real Time Occluded Facial images

In the evaluation of age prediction on non-superimposed occluded images, it was observed that the VGG16 variation model outperformed the ResNet model. The VGG16 model demonstrated a higher accuracy and lower mean absolute error (MAE) in predicting age, indicating its superiority in this task.

6. Future Work

Techniques to enhance occlusion detection and segmentation in face demographic analysis models, to automatically detect and segment occluded regions in facial images, enabling the models to focus on the visible parts of the face for more accurate predictions and improved handling of occlusions can be further implemented. By exploring advanced computer vision methods, such as instance segmentation or generative adversarial networks (**GANs**) these techniques can be successfully implemented. The **FAN** model can also be utilized to improve the project's face detection component.

Additionally, further enhancement of the models' performance, especially in predicting ethnicity is possible by exploring additional architectural variations, **dataset augmentation** techniques, or incorporating other sources of information, such as facial landmarks or contextual cues. Other facial attributes can be explored as well.

7. Conclusion

We addressed the challenge of accurately predicting gender, age, and ethnicity from occluded facial images by proposing three custom CNN architectures based on AlexNet, ResNet, and VGG16. Our models were tailored for multi-task deep learning and demonstrated improved performance compared to existing models. The AlexNet variation model showed limitations in predicting ethnicity and age, leading us to propose an enhanced architecture based on ResNet. This ResNet variation model, with retraining techniques, achieved superior performance for all three attributes compared to the state-of-the-art model. To further enhance the precision in predicting ethnicity, we proposed a VGG16 variation model specifically designed for multi-task learning. This model achieved the highest accuracy rates for gender, ethnicity, and age prediction, surpassing the state-of-the-art model's performance. We utilized the UTK Face Dataset, which provides a diverse range of facial images with annotations for age, gender, and ethnicity. To simulate real-world occlusions, we employed the MaskTheFace tool to generate masked images with various textures, patterns, and colors. The VGG16 variation model achieved high accuracy rates for gender, ethnicity, and age prediction, both in occluded and non-occluded facial images, indicating its robustness and generalization capabilities.

The findings of our research have significant implications in surveillance, marketing, and other real-world applications where face detection and recognition are crucial, especially in environments with frequent facial occlusion. Predicting demographic attributes from occluded images can aid security personnel in identifying potential threats and optimizing their response strategies. Additionally, businesses can leverage this information to deliver targeted ads and promotions that resonate with specific demographics, resulting in enhanced customer engagement and satisfaction. These applications highlight the practical value and potential impact of our model in diverse fields.

References

[1] Prerana Mukherjee, Vinay Kaushik, Ronak Gupta, Ritika Jha, Daneshwari Kankanwadi, and Brejesh Lall: Attribute Prediction in Masked facial images with deep multitask learning: 9th International Conference on Pattern Recognition and Machine Intelligence (PReMI 2021).

[2] Ruder, S. *An overview of multi-task learning in deep neural networks.* arXiv preprint arXiv:1706.05098 (2017).

[3] Wu, G., Tao, J., and Xu, X. *Occluded Face Recognition Based on the Deep Learning.* In 2019 Chinese Control and Decision Conference (CCDC) (pp. 793-797). Nanchang, China: IEEE. doi: 10.1109/CCDC.2019.8832330

[4] Sheoran, V. and Joshi, S. *Age and gender Prediction using Deep CNNs and transfer learning.*

[5] Abrar H. Abdulnabi, Gang Wang, Jiwen Lu, & Kui Jia (2015). Multi-Task CNN Model for Attribute Prediction. IEEE Transactions on Multimedia, 17(11), 1949–1959.

[6] https://towardsdatascience.com/masktheface-cv-based-tool-to-mask-face-dataset-1a71d5b68703

[7] Jianfeng Wang, Ye Yuan, & Gang Yu (2017). Face Attention Network: An Effective Face Detector for the Occluded Faces. CoRR, abs/1711.07246.

[8] Doudou Gao, Peijiang Yuan, Ning Sun, Xulei Wu, & Ying Cai (2017). Face attribute prediction with convolutional neural networks. 2017 IEEE International Conference on Robotics and Biomimetics (ROBIO), 1294-1299.

[9] Yan, X., Yang, J., Sohn, K., Lee, H.: Attribute2image: Conditional image generation from visual attributes. In: European Conference on Computer Vision. pp. 776–791. Springer (2016)

[10] R. Min, A. Hadid and J. -L. Dugelay, "Improving the recognition of faces occluded by facial accessories," 2011 IEEE International Conference on Automatic Face Gesture Recognition (FG), Santa Barbara, CA, USA, 2011, pp. 442-447, doi: 10.1109/FG.2011.5771439.