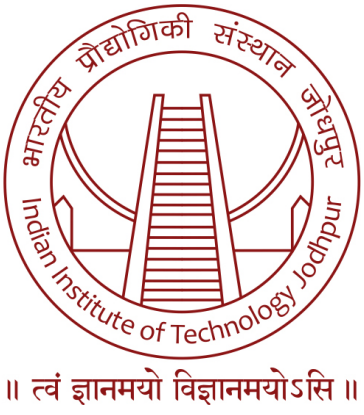
Bias and Fairness in Low Resolution(Very) Image Recognition

A Project Report Submitted by

Sasikanth Kotti

in partial fulfillment of the requirements for the award of the degree of

M.Tech



Indian Institute of Technology Jodhpur

Computer Science and Engineering

January, 2023

Declaration

I hereby declare that the work presented in this Project Report titled Bias and Fairness in Low Resolution(Very) Image Recognition submitted to the Indian Institute of Technology Jodhpur in partial fulfilment of the requirements for the award of the degree of M.Tech, is a bonafide record of the research work carried out under the supervision of Dr. Mayank Vatsa and Dr. Richa Singh. The contents of this Project Report in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

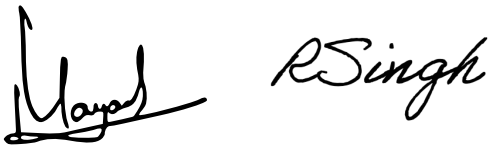
Signature

Sasikanth Kotti

MT19AIE308

Certificate

This is to certify that the Project Report titled Bias and Fairness in Low Resolution(Very) Image Recognition, submitted by Sasikanth Kotti(MT19AIE308) to the Indian Institute of Technology Jodhpur for the award of the degree of M.Tech, is a bonafide record of the research work done by him under my supervision. To the best of my knowledge, the contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.



Signature

Dr. Mayank Vatsa and Dr. Richa Singh

Acknowledgements

I would like to thank my supervisors **Dr. Mayank Vatsa** and **Dr. Richa Singh** for their continuous support and precious guidance without which this work would not have been possible. They had inculcated within me enthusiasm and love for research. This inspired me to pursue research for the rest of my life to the extent possible.

Additionally I want to extend my thanks to lab mates from Image Analysis and Biometrics (IAB) Lab particularly **Kartik Thakral** and **Surbhi Mittal** for their insightful and useful discussions. I want to extent my gratitude towards all the faculty members of my department for imparting knowledge and enabling joyous learning.

I would like to mention **Kishore Kodipaka**, **Sukarna K**, **Vidyaranya Gujju** and **Sri C.K. Prasad** for inspiring me with this journey. I would also like to thank my parents, friends, batch mates and family members for their constant encouragement and support. Lastly I want to mention my son **Jayaditya Naga Sai** and my wife **Vani** for not demanding time which helped me to carryout this important work of my lifetime....

Abstract

Recent image recognition algorithms showed great performance, which was possible due to advancements in deep learning methods. These methods find application in variety of scenarios such as recognising species, surveillance, identifying missing persons, identifying objects from drones during floods etc. However, current methods assume the availability of images of very high resolution. Obtaining images of high resolution is not always possible due to large distance of the object and inherent limitations in the acquisition device such as camera. Hence it is important to develop robust algorithms that can also work with low resolution and very low resolution images. It is also essential that these algorithms are not biased and are fair to different sub groups.

Recent algorithms in the literature showed decent improvement in performance. However, bias and fairness was not taken into consideration in the current algorithms specially in components that use generative models.Hence, there is enormous scope to further improve the performance of image recognition with low resolution along with developing fair algorithms. In this work, we attempted to cover existing literature. We showed experimentally that this problem needs further research for improved performance. We had also showed that existing generative models specifically GANs are prone to bias and fairness issues and can cause cause disparate impact. Lastly we proposed methods and techniques to debiase existing generative models. We hope these techniques can be used to develop fair algorithms for low resolution image recognition.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Biological vision is the dominant method by which most of the living beings and humans perceive the environment. Inspired by this phenomena, computer scientists attempted to enable vision for computers. Although not exactly same computer vision consists of problems related to variety of tasks such as recognition, classification, detection and segmentation. The dominant task among these is image recognition which has wide applications for real world problems. Of these face recognition is of specific interest to research community. Some of the prominent applications of face recognition are :

- Surveillance
- Biometric security
- Identity verification
- Law enforcement

There are many covariates of face recognition such as recognising faces with variation in illumination, occlusion, impersonation etc. All of these assume the availability of high resolution face images. However, in scenarios such as long distance surveillance camera settings or recognising from drones high resolution image of faces are not always available. High resolution faces may also be not available due to the limitations of the camera specifications especially in settings such as previous generation of mobile devices. To perform face recognition with low resolution images normal interpolation of images to the desired resolution is not helpful due to information loss. Many algorithms [1] [2] were developed to solve this challenging problem.

Most of these algorithms and architectures for low resolution face recognition utilize generative models as one of the major components. This component helps in enriching the representations of low resolution images. However, as illustrated in figure 1.0.1, generative models can enrich the representations towards any sub group. This can result in disparities in down stream applications. Considering the criticality of the applications in which low resolution face recognition algorithms are deployed it is essential that these algorithms are bias free and doesn't disparately effect different sub groups.

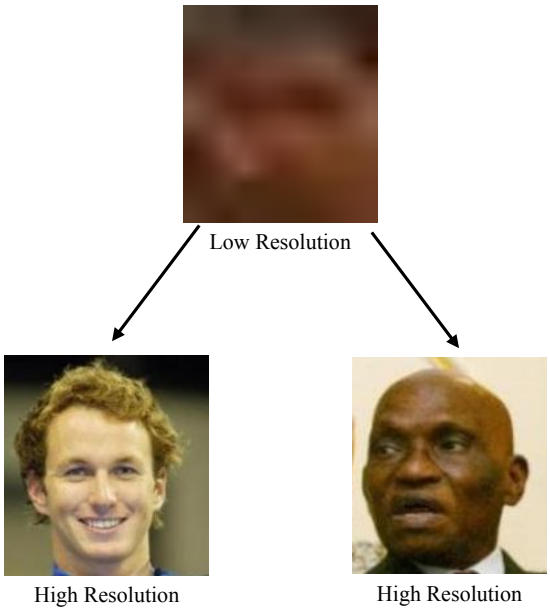


Figure 1.0.1: Low Resolution, High Resolution images for given subject

Hence, this problem of fairness and bias in generative models for low resolution face recognition is an important and open problem because of the challenges and the real world implications. Hence, development of fair algorithms for face recognition across covariates can increase trust and effective usage.

1.1 Literature Survey

1.1.1 Constrained Face Recognition

There are two broadly two category of unconstrained face recognition methods proposed in the literature

Shallow Face Recognition

These are the foundational methods initially proposed for face recognition.

- **Face Recognition Using Eigenfaces** [3] is a foundational method proposed by Turk et.al where faces are projected into feature space defined by eigen vectors. These are called as eigen faces. These features are compared with the features of the known faces for detection and recognition.
- **Eigenfaces vs. Fisherfaces: recognition using class specific linear projection** [4] is yet another foundational method proposed by Belhumeur et.al. Here the authors project faces into subspace so that the deviations are minimal. This projection is based on Fisher's Linear Discriminant which gave better performance and lower error rates than that of eigen faces method.

Deep Face Recognition

- **LightCNN29** [5] The idea of this architecture proposed by Wu et.al is to define a Light CNN framework to learn representations even from large scale noisy data. In this framework a variation of maxout activation, called Max-Feature-Map (MFM) is introduced along with each convolution layer.
- **VGGFace2** [6] A large scale face dataset with variations across pose and age was introduced by Cao et.al. Improved performance on face recognition was demonstrated by training standard architectures such as Resnet50 on this dataset.
- **SphereFace** [7] Liu et.al proposed the idea of enabling convolution neural networks to learn angular discriminative feature for face recognition along the hypersphere.
- **Additive Margin Softmax** [8] In this method a variant of softmax loss is proposed by Wang et.al where the angular margin is incorporated additively resulting in more interpretable loss function. Also this resulted in better class separability.
- **CosFace** [9] Hao Wang et.al proposed this method which consists of reformulating softmax loss as cosine loss and maximizing the margin in angular space.
- **ARCFace** [10] In this method a novel loss function that incorporates margins is proposed by Deng atl.al. This proposed loss has clear geometric interpretation where different classes are well separated along the hypersphere.

- **CurricularFace** [11] In this method Huang et.al introduces curriculum learning for face recognition. This is achieved by introducing a modulation term in existing loss functions to incorporate importance of easy and hard samples during training.

- **DeepFace-EMD** [12] Proposed by Phan et.al this method employs Earth Mover's Distance on the deep, spatial features of image patches for Face identification. This approach helps even the pretrained model generalize better to OOD face images.

1.1.2 Low Resolution Face Recognition

Image enhacement Based

- **Face Hallucination Using Cascaded Super-Resolution and Identity Priors** [13], In this method proposed by Grm et.al low-resolution images are upsampled using convolution neural network which are guided by face recognition models which are ensembled to act as identity priors

- **SiGAN: Siamese Generative Adversarial Network for Identity-Preserving Face Hallucination** [14], proposed by Hsu et.al this method utilizes generative modelling where generated faces are guided by identities. This preserves visual information corresponding to identities ,thereby helping with recognition.

- **SUPREAR-NET: Supervised Resolution Enhancement and Recognition Network** [15], Ghosh et.al proposed Supervised Resolution Enhancement and Recognition Network based on generative adversarial networks.This transforms a low resolution probe into high resolution without corrupting useful class specific information.

- **Face hallucination using convolutional neural networks** [16], the authors in this **patent** proposed an approach to generate higher resolution face image being called as hallunicated face image by using a bi-channel deep convolutional neural network (BCNN).

Classifier Based

- **DeriveNet for (Very) Low Resolution Image Classification** [1], this current state of the model by Singh et.al proposes a novel DeriveNet model for LR/VLR recognition. This is achieved by couple of novel loss formulations namely "Derived-Margin softmax loss" and "Reconstruction-Center (ReCent) loss".

- **Dual Directed Capsule Network for Very Low Resolution Image Recognition** [2], Singh et.al in this algorithm proposed novel architecture named DirectCapsNet along with "HR-anchor loss" and "targeted reconstruction loss" for LR/VLR recognition. These loss functions help overcome limited information content in LR/VLR images.

- **Enhancing Fine-Grained Classification for Low Resolution Images** [17], Singh et.al proposed a novel loss function namely attribute-assisted loss that utilizes ancillary information for

classification.

- **Improved Knowledge Distillation for Training Fast Low Resolution Face Recognition Model** [18], Wang et.al proposed a variation of knowledge distillation to train student with LR augmented data. The distributional discrepancy is addressed by constraining the multikernel maximum mean discrepancy between outputs of student and teacher.

- **Exploring Factors for Improving Low Resolution Face Recognition** [19] The authors Omid Abdollahi Aghdam et.al investigated factors such as variation of appearance , resolution distribution for training data and importantly resolution difference between images of gallery and probe.

- **Human facial feature detection method under low resolution** [20], in this **patent** the authors proposed to detect low resolution face features by using integral computation and integrogram.The face from the video is detected using HAAR based ADABOOST algorithm. After the face is segmented to detect features related to mouth and nose.

1.1.3 Bias and Fairness

Bias and Face Recognition

- **Deep learning for face recognition: Pride or prejudiced?** [21], The authors Nagpal et.al in this work provided comprehensive analysis of bias in deep learning based facial recognition systems and algorithms.

- **Face Recognition: Too Bias, or Not Too Bias?** [22], Authors Robinson et.al in this work provides insights for bias in current state of the art face recognition systems by proposing BFW dataset. The inherent bias in humans is also evaluated.

- **Consistent instance false positive improves fairness in face recognition** [23], The authors Xu et.al in this work proposed "false positive rate penalty loss" to mitigate bias in face recognition. As a result , this method mitigates bias without the need for demographic annotations.

Bias and Generative Models

- **Fair attribute classification through latent space de-biasing** [24], Ramaswamy et.al introduced a method for training accurate attribute classifiers using GAN augmented data. The inherent biases in GAN are removed from the augmented data by perturbing the underlying latent space.

- **Jointly de-biasing face recognition and demographic attribute estimation** [25], A novel debiasing adversarial network (DebFace) is proposed by Gong et.al to extract disentangled feature representations for unbiased face recognition and demographics estimation.

- **FairStyle: Debiasing StyleGAN2 with Style Channel Manipulations** [26], Karakas et.al demonstrated in this method that the style space of StyleGAN2 model is used to perform disen-

tangled control of the target attributes for debiasing.This is done without training any additional models.

- **Imperfect ImaGANation: Implications of GANs Exacerbating Biases on Facial Data Augmentation and Snapchat Selfie Lenses** [27], Jain et.a in this work demonstrates how popular GANs exacerbate biases specifically with respect to gender and skin tone for faces.

1.1.4 Bias and Distillation

Generative Models

- **Distilling the knowledge in a neural network (2015)** [28], This is the first work in which the authors Hinton et.al introduced knowledge distillation. Authors performed distillation by compressing the knowledge of large ensemble of models into single simpler model. This is done via softmax probabilities.

- **Compressing gans using knowledge distillation** [29], Aguinaldo et.al explored distillation from large teacher GAN to student GAN. This is achieved by using MSE loss and Joint loss function that supervises regular GAN training with MSE loss.

- **Online multi-granularity distillation for gan compression** [30], In this work Ren et.al proposed a novel online multi-granularity distillation (OMGD) scheme to obtain lightweight GANs. In this online approach, progressively promoted teacher GANs refined student generators which are disriminator free.

- **Tinygan: Distilling biggan for conditional image generation** [31], The authors Chang et.al proposed distillation for GANs without access to internals of the model. This is achieved by proposing novel loss formulations and distillation process.

Classification Models

- **Fairness via Representation Neutralization** [32], In this work the Du et.al demonstrated that fairness can be improved by just debiasing the classification head of DNN models. This was done by introducing a mitigation technique named "Representation Neutralization for Fairness (RNF)".

- **One-network adversarial fairness** [33], A fair adversarial discriminative model is proposed by Adel et.al to achieve fairness. This was proposed as a plug and play component so that a potentially unfair deep architecture can be trained with fairness contraints.

- **Constructing a fair classifier with generated fair data** [34], Jang et.al proposed to train the model by generating synthetic data which is fair w.r.t multiple sensitive attributes. The classifier is then transferred to real data.

- **Fair classification with adversarial perturbations** [35], In this work Celis et.al studied fair classification where protected attributes of a fraction of samples are arbitrarily perturbed. They

proposed an optimization framework for learning fair classifiers in this setting with provable guarantees.

1.2 Research Motivation

The task of Image recognition consists of identifying and classifying objects in scenes. Although recent techniques in computer vision achieved impressive performance in object recognition and classification, their performance degrades considerably with low resolution images. Additionally, existing techniques may be biased and unfair to different sub groups.This is relevant in real world settings where it is not practical and always not possible to obtain high resolution images. This is due to variety of environmental conditions and location of the camera. Also low resolution image recognition plays an important role in applications such as surveillance, arial survey during floods, identifying objects of interest from long distance among others.

1.3 Research Contributions

Generative models such as GANs can aid in improving the performance of low resolution face recognition systems. These enable enhacing and enriching the representations of low resolution images in comparison to high resolution images.Hence, understanding and improving the characteristics of these generative models can enable low resolution face recognition systems to perform equally for all sub groups. In this thesis, we demonstrate :

- The performance of current face recognition systems such as LightCNN29 [5], ArcFace [10] and VGGFace2 [6] degrade despite fine-tuning with low resolution face images.
- Pretrained GANs such as StyleGAN2 with adaptive discriminator augmentation (ADA) [41] are biased towards specific sub groups for different attributes such as race, race4 and age.
- Using synthetic data from pretrained GANs such as StyleGAN2 with adaptive discriminator augmentation (ADA) [41] for fine tuning can disparately impact performance among sub groups in face verification systems (VGGFace2 [6]).
- Proposed methods and techniques such as *Fair Sampling* and *Shannon Diversity Loss* to mitigate bias for the student GAN during knowledge distillation from pretrained GANs such as StyleGAN2 with adaptive discriminator augmentation (ADA) [41]

Chapter 2

Proposed Algorithm

This chapter discusses the research methodology, research findings and proposed algorithm. A detailed overview of the overall findings and algorithm is presented in different sections.

2.1 Image Recognition Performance

This section describes the methodology for understanding the peformance of image recognition specifically face recognition tasks. This is obtained by performing domain adaptation of pretrained models using LFW [36] dataset. We also show the performance of domain adaptation on low resolution images using QMUL-SurvFace [37] dataset. We briefly describe the datasets, models, implementation details and protocols followed for training and evaluation.

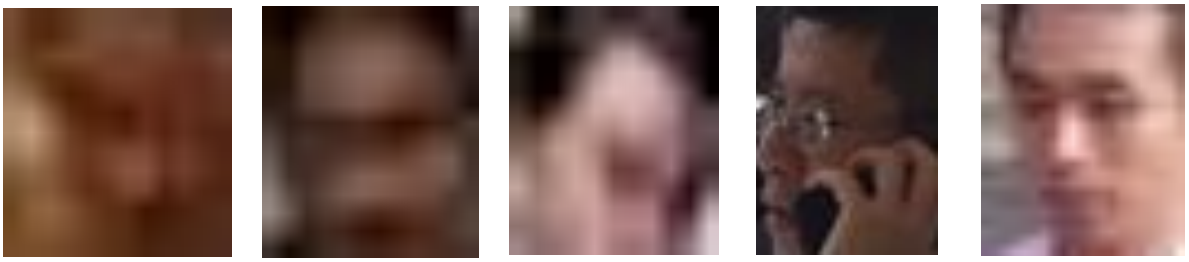
2.1.1 Face Recognition Datasets

- **LFW dataset** i.e. Labeled Faces in the Wild is a public benchmark that can be used for face recognition and verification. LFW deep funneled images [36] consists of LFW images aligned using deep funneling as shown in (a) figure 2.1.1. This dataset consists of 13233 aligned face images. This has 5749 persons of which 1680 identities have two or more images. The protocol for face verification consists of splitting the dataset into train and test of which train has 1180 identities and test has 500 identities. The test set consists of probe and gallery with non-overlapping face images.

- **QMUL-SurvFace** [37] consists of native low-resolution face images. These are not synthesised by artificial down-sampling of high-resolution face images shown in (b) figure 2.1.1. There are 463,507 face images belonging to 15,573 identities which are distinct and captured in real-world scenarios.The trainset contains 220,890 face images belonging to 5,319 identities and testset contains 242,617 face images belonging to 10,254 identities. The gallery in testset contains 60,294 images from 5,319 test identities whereas mated probe contains 60,423 probe images from 5,319 test identities.

The remaining 12,1736 distractor probe face images without mated gallery are available in unmated probe. Only mated probe images are used for evaluation.

(a) LFW Dataset



(b) QMUL-SurvFace Dataset

Figure 2.1.1: LFW and QMUL-SurvFace Datasets

2.1.2 Pre-trained Models

- **LightCNN29** [5] is a 29 layer CNN model, where a variation of maxout activation known as Max-Feature-Map (MFM) is introduced in each convolution layer. The model trained with large scale noisy face datasets such as CASIA-WebFace and MS-Celeb-1M in gray-scale is used as pre-trained model. This model has a feature dimension of 256.

- **ArcFace** [10] is a resnet18 backbone pre-trained with MS1MV3 dataset with Arcface i.e. Additive angular margin loss for deep face recognition. The feature dimension of the embedding for this model is 512.

- **VGGFace2** [6] is a resnet50 backbone trained with MS-Celeb-1M and the fine-tuned with VGGFace2 dataset. This pre-trained model has a feature dimension of 2048.

2.1.3 Implementation Details

We describe below the implementation details and preprocessing steps for fine-tuning face recognition models.

The different steps before fine-tuning as shown in figure 2.1.2 are described below:

Figure 2.1.2: Face Recognition Pre-processing Pipeline

- **Face Detection and Cropping**, images that are used for face recognition can consists of variety of objects in addition to faces. Inorder to fine-tune and use face recognition models, it is essential to seperate out the faces from other objects. This is performed by initially detecting the faces and obtaining the coordinates of the bounding boxes for the faces. Face detection can be performed by using dlib, MTCNN face detectors. These coordinates are then used to just crop the faces from the image. The resultant face image then contains only face features which can be used for training or inference from a face recognition model. Alternatively, if the dataset is contrained, clean and contains single face in each image a direct center crop can also result in extracting the face. A center crop of 128x128 and 112x112 is used for LFW dataset when fine-tuning VGGFace2 and ArcFace models. Random Crop of 128x128 is used for LFW dataset when fine-tuning LightCNN29 model. Face detection and cropping is not done for QMUL-SurvFace dataset. Faces cannot be recognized in QMUL-SurvFace dataset due to very low resolution images.

- **Face Alignment and Resizing**, although faces can be extracted from the images these may be not in the correct alignment. This can be specifically observed when the images are based out of different actions instead of just portraits. Hence, the extracted face need to aligned. The alignment is carried out by identifying the salient points in the face and then moving face. Face recognition models require input images to be of specific size such as . Hence the aligned face images are resized before feeding to the network. The LFW dataset version used consists of already pre-aligned images with deep funneling approach. Hence, no additional alignment is performed. Alignment is also not performed on faces from QMUL-SurvFace due to very low resolution. The face images from both the dataset are resized to 128x128, 224x224 and 112x112 before feeding to LightCNN29, VGGFace2

and Arcface models respectively.

- **Data Augmentation**, images are augmented with additional transformations such as RandomHorizontal flip, Random Grayscale and RandomCrop. Augmentation artificially increases the number of training samples and helps model to generalize better during inference. It is to be noted that images are converted to grayscale before feeding to LightCNN29 pre-trained network. However, for other networks i.e. VGGFace2 , Arcface RGB images are used for fine-tuning.

- **Feature Extraction**, variety of features such as Histogram of gradients (HOG), Scale-Invariant feature Transform (SIFT) and Local Binary Patterns (LBP) can be extracted from face images for querying the gallery images with probe during inference. However, these hand-crafted features doesn't provide good performance. However, deep features from fine-tuned networks can provide better performance. We had extracted these deep features after fine-tuning LightCNN29, VGGFace2 and Arcface networks. These features are 256, 512 and 2048 dimensional embeddings respectively.

- **Fine-tuning**, Pre-trained LightCNN29 [5], ArcFace [10] and VGGFace2 [6] models are considered for domain adaptation with LFW [36] dataset and QMUL-SurvFace [37] as per the protocol. These models are fine-tuned by freezing the last 10 layers with contrastive loss [38] .

  Fine-tuning with LFW was carried out for all the models, for 50 epochs with a learning rate of 1e-5 and with batch size of 128. Weight decay of 1e-2 was used for ArcFace [10] and VGGFace2 [6], where as LightCNN29 [5] was fine-tuned with weight decay of 1e-4.

  Fine-tuning with QMUL-SurvFace [37] dataset was carried for LightCNN29 [5], ArcFace [10] and VGGFace2 [6], for 50 epochs with a learning rate of 1e-2 , batch size of 1024,128, 128 and weight decay of 1e-4 respectively.

2.1.4 Protocols and Evaluation

Inference is carried out by using cosine distance between probe and gallery image embeddings. For each probe image, the embedding vector is generated from the fine-tuned model. Similarly embedding vectors are also generated for all the gallery images. Cosine distance is then computed between the probe embedding and all the gallery embeddings.

Rank1 and Rank10 identification accuracies were obtained for both these datasets for each of the fine-tuned models.

CMC curves were plotted using identification accuracies on y-axis and rank on x-axis. These CMC curves are used for anlysing the performance of the fine-tuned models for both datasets i.e. LFW and QMUL-SurvFace.

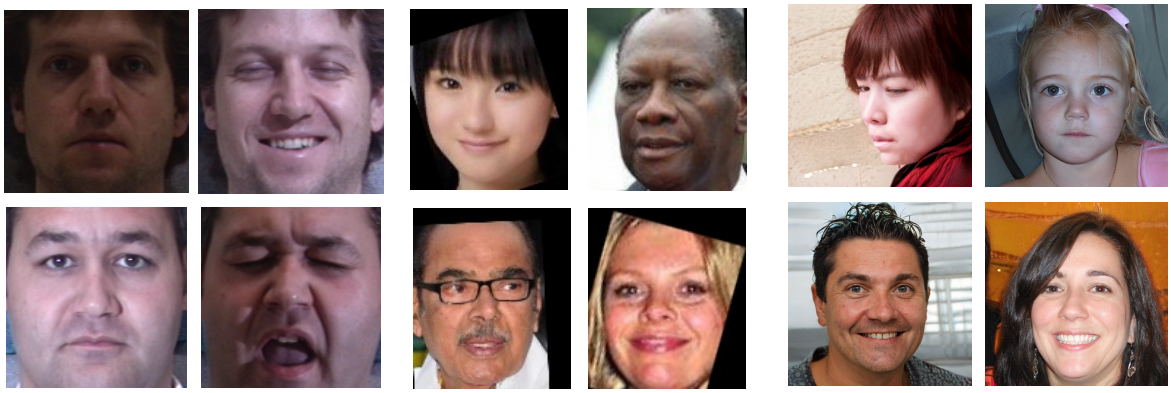
2.2 Bias and Fairness of GANs

As part of this work, bias and fairness of existing generative models are understood by obtaining attributes and performing domain adaptation on pretrained models. This is specifically understood in the context

of face verification task. We describe below different datasets and protocol followed for evaluating fairness and bias.

2.2.1 Face Verification Datasets

- **CMU Multi-PIE** [39] is a constrained dataset consisting of face images of 337 subjects with variation in pose, illumination and expressions. Of these over 44K images of 336 subjects images are selected corresponding to frontal face images having illumination and expression variations as shown in (a) figure 2.2.1
- **The Balanced Faces in the Wild (BFW) dataset** [22] , is balanced across eight subgroups This consists of 800 face images of 100 subjects, each with 25 face samples.The BFW dataset is grouped into ethnicities (i.e., Asian (A), Black (B), Indian (I), and White (W)) and genders (i.e., Females (F) and Males (M)) shown in (b) figure 2.2.1. The metadata for this dataset consists of list of pairs for face verification. Hence, this dataset can be used to investigate bias in automatic facial recognition (FR) system for verification protocol.
- **FFHQ** which stands for Flickr-Faces-HQ [40] shown in (c) figure 2.2.1 is a dataset of 70,000 human faces of high resolution 1024x1024 and covers considerable diversity and variation.



(a) CMU Multi-PIE (b) BFW (c) FFHQ

Figure 2.2.1: CMU Multi-PIE, BFW and FFHQ Datasets

Evaluation for estimation of bias and fairness is performed in two phases. Initially, the proportion of faces generated for each sub-group of different attributes such as Age, Gender, Race and Race4 were analysed. In the next phase, a pretrained face verification model is fine-tuned, and the impact of fairness

is analyzed using Degree Of Bias(DoB) metric. We define DoB for face verification as the standard deviation of GAR@FAR

$$DoB_{fv} = \sqrt{\frac{\sum (GAR_{sg} - \mu)^2}{N}} \tag{2.1}$$

where GAR_{sg} stands for GAR @ FAR for each sub-group, μ represents mean of GAR@FAR and N represents number of sub-groups. The GAR and FAR stands for Genuine Accept Rate and False Accept Rate respectively.

2.2.2 Pre-trained Models

We briefly describe the pre-trained models used to understand and evaluate bias and fairness in GANs.

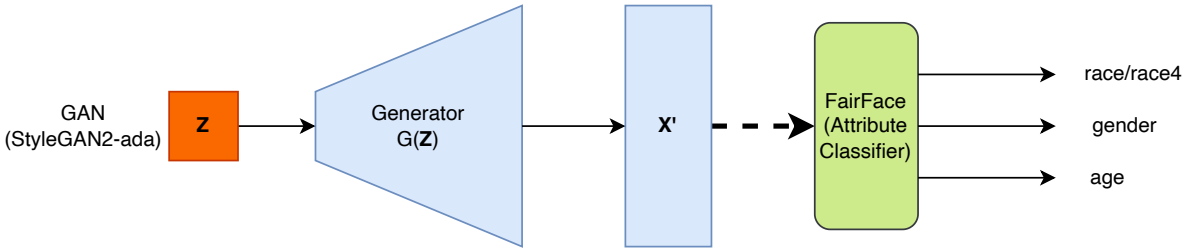
- **Fairface Classifer**, is a resnet34 model trained on FairFace dataset. As part of the training, faces are detected, cropped and aligned using dlib models. The face images are resized to 224x224 and then normalized for training. The same preprocessing steps are also used during inference to obtain race, gender and age attributes.
- **VGGFace2**i, s a resnet50 backbone trained with MS-Celeb-1M and then fine-tuned with VGGFace2 dataset.This pre-trained model has a feature dimension of 2048.
- **StyleGAN2-ADA**, is StyleGAN2 architecture that was trained using adapative discriminator augmentation mechanism. Model trained with FFHQ dataset is used as pre-trained model for analysis.
- **DiscoFaceGAN**, is a pretrained model where faces of non-existent people with variations of pose, expression and illumination can be generated. The model is trained using imitative-contrastive learning to learn dientangled representations. This model trained with FFHQ data set is considered for analysis.

2.2.3 Implementation Details

Experiment-1: As part of this experiment the generator of StyleGAN2 with adaptive discriminator augmentation (ADA) [41] trained on FFHQ dataset is used to generate synthetic face images. Attributes such as race, race4, gender and age of these synthetic faces were obtained using a pretrained Fairface [42] attribute classifier as illustrated in figure 2.2.2. The proportion of images for each attribute type were plotted as shown in figure 3.3.2, to understand bias and imbalance

Experiment-2: In this experiment DiscoFaceGAN [43] is considered for generating different faces for different identities, expressions, lightning and poses. These are shown in figure 2.2.3.

VGGFace2 [6] model is considered for domain adaptation wth CMU Multi-PIE [39] and synthetic faces generated with DiscoFaceGAN [43]. About 10000 synthetic faces of 2500 identities as shown in figure 2.2.4 were generated with with DiscoFaceGAN [43]. Out of these, 2000 identities were used for



FairFace and StyleGAN2

Figure 2.2.2: GAN Bias Estimation Architecture

GAN Generated Faces

Figure 2.2.3: DiscoFaceGAN Generated Faces

training and 500 identities were used for validation. The 336 subjects of CMU Multi-PIE [39] were split into 70-30 ratio for training and validation. Fine-tuning was carried out for 10 epochs with a learning rate of 1e-4 , batch size of 128, weight decay of 1e-4 and momentum of 0.9. The last two convolutional layers of VGGFace2 [6] were fine-tuned with ArcFace [10] loss of margin 35 and scale 64. The checkpoint with lowest validation loss is considered for inference with BFW dataset [22].

2.2.4 Evaluation Protocols

Inference is carried out by using Cosine distance between the pairs of BFW dataset.This entire pipeline is illustrated in figure 2.2.5. Comparison of DoB_{fv} i.e Std(GAR @ FAR) for different attributes such as race, gender and others is carried out for both models i.e. models fine-tuned with CMU Multi-PIE and Synthetic Faces.

Synthetic(DiscoFaceGAN) Faces

Figure 2.2.4: Synthetic DiscoGAN Faces

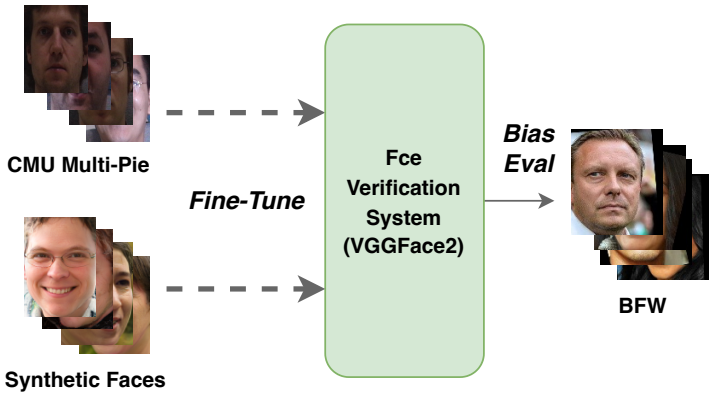


Figure 2.2.5: Bias Estimation in Face Verification System

2.3 Fairness in Distillation of GANs

2.3.1 Algorithm Description

Distillation in GAN focusses on transferring the knowledge and representations from a large teacher generator to small student generator. Ting-Yun Chang et.al proposed black box distillation framework for compressing GANs [31]. In this work, the authors generated images from the teacher GAN and then trained the student GAN with these generated images by using a combination of Pixel-Level Distillation Loss [31] , Adversarial Distillation Loss [31] and Feature-level Distillation Loss [31].

We built on top of this existing work TinyGAN [31] and proposed fair distillation to obtain fair student GAN. The architecture of fair distillation for GANs is shown in figure 2.3.1.

Below are our contributions:

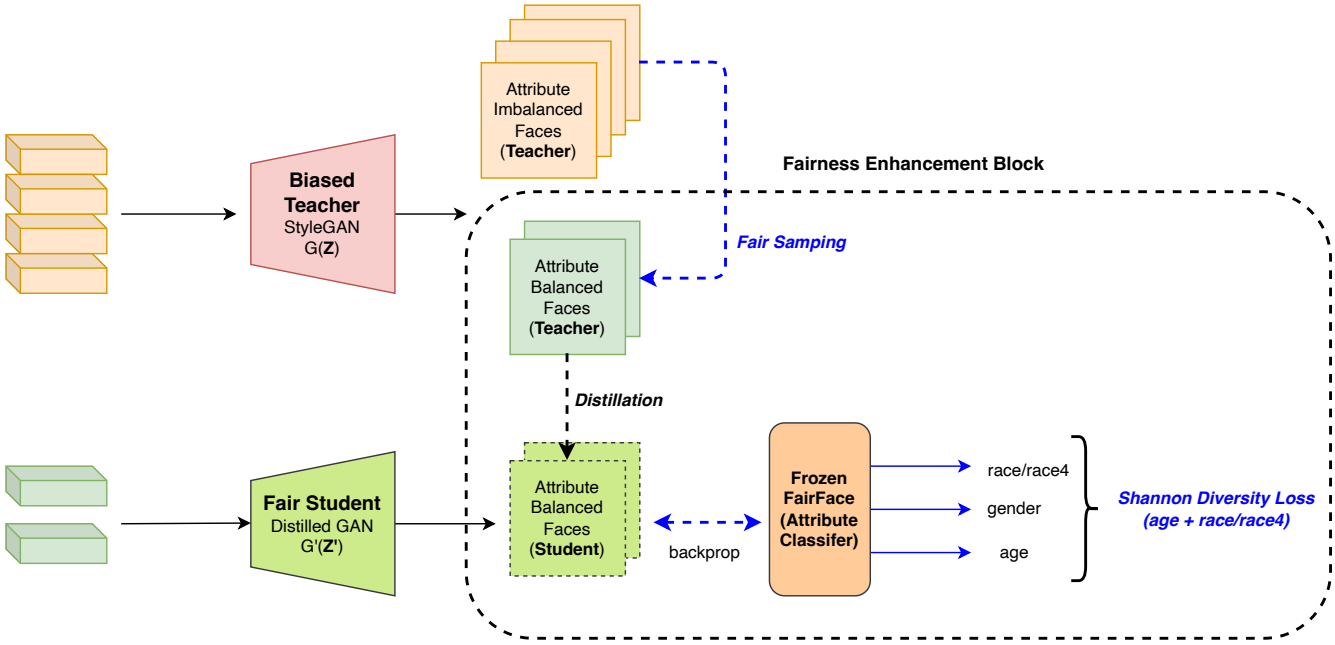


Figure 2.3.1: Fair Distillation Architecture

- **Fair Sampling** , A fairface attribute classifier is used to obtain the attributes of all face images generated by the teacher GAN. Uniform number of face images are sampled for each category of the attribute (race/race4, gender and age) classified by FairFace classifer.
- **Shannon Diversity Loss** , is derived from Shannon index which is used to measure diversity in ecology. We propose to use Shannon Diversity Loss from categories of each attribute (race/race4 and age) of FairFace classifier as an additional fairness contraint during the distillation process.

We show below losses as defined by Ting-Yun Chang et.al [31] in addition to the Shannon Diversity Loss.

Pixel-Level Distillation Loss [31] :

$$L_{\text{KD\_pix}} = E_{z\sim p(z))}[\|T_G(z) - S_G(z)\|_1], \tag{2.2}$$

where T_G is the frozen teacher network (StyleGAN generator), S_G is student network, $z \in R^{512}$ is a latent variable

Adversarial Distillation Loss [31] :

$$L_{\text{KD\_S_G}} = -E_z[D(S_G(z))] \tag{2.3}$$

for the generator, and the loss

$$L_{\text{KD\_D}} = E_z[\max(0, 1 - D(T_G(z))) + \max(0, 1 + D(S_G(z)))] \tag{2.4}$$

for the discriminator, where z is the noise vector, $T_G(z)$ is the image generated by StyleGAN, while S_G and D are respectively the generator and discriminator representing Student

Feature-Level Distillation Loss [31] :

$$L_{\text{KD\_feat}} = E_z[\Sigma_i \alpha_i \| D_i(T_G(z)) - D_i(S_G(z)) \|_1], \tag{2.5}$$

Shanon Diversity Loss :

$$L_{\text{SD}} = \sum_{attribs} (1 - \frac{\Sigma_i{}^k p_i log(p_i)}{log(k)}) \tag{2.6}$$

where p_i represents proportions and k total groups or entities for each attribute. This loss is computed for each attribute (race/race4 and age) and aggregated together.

Overall Loss : The overall loss to perform distillation from teacher GAN to student GAN is now defined as below.

$$L_{\text{S}} = \lambda_1 L_{\text{KD\_feat}} + \lambda_2 L_{\text{KD\_pix}} + \lambda_3 L_{\text{KD\_S}} + \lambda_4 L_{\text{SD}}, and \tag{2.7}$$

$$L_{\text{D}} = L_{\text{KD\_D}} \tag{2.8}$$

2.3.2 Implementation Details

Initially 80,000 images were generated by the teacher GAN (StyleGAN). Of these around 18,329 images were sampled based on the attributes from fairface classifier. These sampled images are then used to perform distillation of the student GAN. We had used the same architecture for student generator and student disriminator as used by [31]

Distillation with fairsampled images was carried out with student generator and student discriminator for 250 epochs with a batch size of 128. An initial learning rate of 1e-3 is used for the training. Adam optimiser with beta1, beta2 as 0.0,0.9 , weight decay of 1e-4 and CosineAnnealing schedule is used for optimal results.

We followed the same approach as [31] where the pixel level distillation loss $\lambda_2$is decayed to zero. λ_1 is set to 0.4 and λ_3 is set to 0.3 for feature distillation and adversarial distillation losses respectively.

The shannon diversity lossλ_4 is linearly scaled from 0 to 0.3 during the training process. This based on the fact that the student GAN doesn't generate clear faces during initial training process. Hence the fairface classifier may not predict accurate attributes. As the training converges the fairface classifier predicts accurate attributes due to clearer images generated by the student GAN. The enables accurate computation of the shannon diversity loss and its contribution as fairness constraint in the loss function.

After training for 250 epochs, generator checkpoint at 201 epoch is used for inference. The checkpoint is selected based on the visual inspection of sample images generated by the student generator.

2.3.3 Evaluation Protocols

5000 images were generated by the distilled student generator. The attributes of these images were obtained by using pretrained fairface classifier. The proportion of total images for each category of attributes (race/rage4, age and gender) were plotted to understand fairness and imbalance.

2.3.4 Ablation Study

To better understand the role of **Shannon Diversity Loss** ablation study was conducted. As part of this, all the images generated by the Teacher GAN were used for distillation. This effectively removes **Fair Sampling** component. Distillation of Student GAN was carried out using the same hyper paramters. Training i.e. distillation was carried out for 200 epochs with batch size of 128 and an initial learning rate of 1e-3 . Adam optimiser with beta1, beta2 as 0.0,0.9 , weight decay of 1e-4 with CosineAnnealing schedule was used during distillation.

The distilled Student GAN was then used to generate 5000 images. The attributes obtained with fairface classifier when plotted as proportion of all images revealed no changes to fairness and imbalance. Further research is needed to understand the role and impact of **Shannon Diversity Loss**. This research is planned for future work.

Chapter 3

Experimental Results

3.1 Results for LR(V) face recognition on QMUL-SurvFace dataset:

CMC Curves of fine-tuning all architectures (LightCNN29, ArcFace and VGGFace2) with LFW and QMUL-SurvFace datasets are displayed in figure 3.1.1 and figure 3.1.2 . From these CMC Curves, it is evident that performance degrades when using low resolution face images QMUL-SurvFace, when compared to high resolution face images LFW dataset. This is across all the ranks and despite fine-tuning all the models with QMUL-SurvFace dataset

Table 3.1.1 shows the results of fine-tuning challenging QMUL-SurvFace dataset and LFW on different networks. On comparing the Identification accuracy at Rank1 and Rank10, it can be observed that the performance on QMUL-SurvFace dataset is worse than that of LFW dataset. Also the current state of the art on QMUL-SurvFace is 72.34% accuracy achieved by DeriveNet [1] model. From these results it can be inferred that face recognition on LR/VLR images is a challenging task that needs further research. It can also be concluded that existing architectures cannot be adopted as-is for face recognition with low-resolution images.

Model	Rank1		Rank10	
Dataset	LFW	QMUL-SurvFace	LFW	QMUL-SurvFace
LightCNN29	96.8	6.3	98.6	15.2
VGGFace2	95.6	1.6	99.7	5.7
ArcFace	80.7	1.8	89.6	5.8

Table 3.1.1: Identification Accuracy (FineTuned with LFW and QMUL-SurvFace).

3.2 Results for bias and fairness of generative models:

Generative models especially GANs are important components in most of the architectures designed to perform low resolution image recognition.

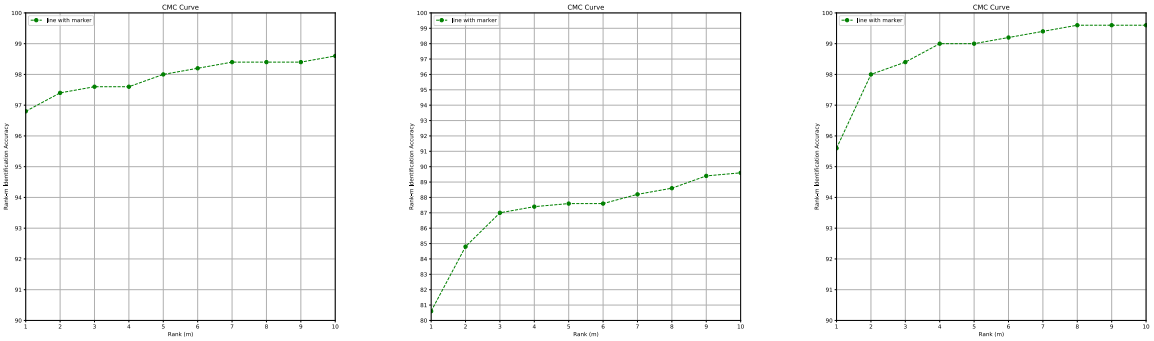


Figure 3.1.1: LightCNN29, ArcFace, VGGFace2 - CMC Curve(LFW)

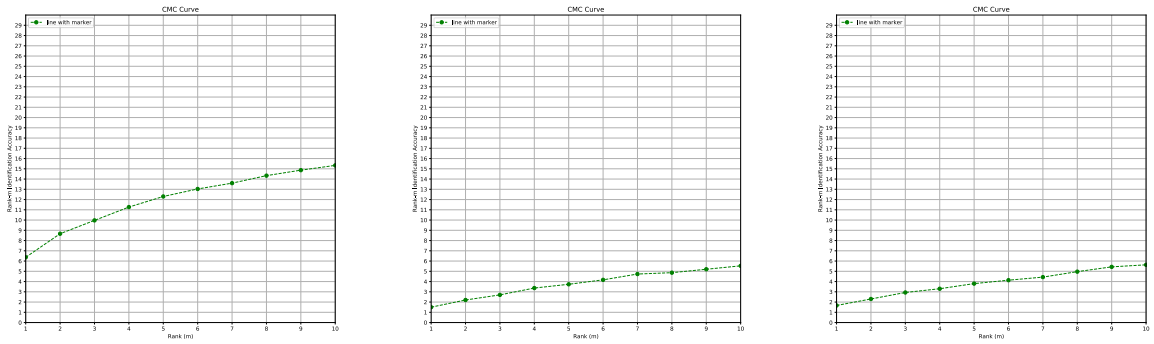


Figure 3.1.2: LightCNN29, ArcFace, VGGFace2 - CMC Curve(QMUL-SurvFace)

To better understand the impact of bias in generative models,a pretrained face recognition model VGGFace2 [6] is fine-tuned with both CMU Multi-PIE [39] and synthetic faces generated with Disco-FaceGAN [43] for face verification. The GAR@FAR performance of both fine-tuned verification models for BFW dataset are shown in Table 3.2.1. GAR@FAR for sub-groups of ethnicity and gender for BFW dataset are shown in Table 3.2.2 and Table 3.2.3.

From figure 3.3.2, it is evident that GANs trained with the FFHQ dataset are biased towards generating more faces in the age group "20-29" and mostly "White" faces. However, no such imbalance is observed for gender attribute.

Figure 3.3.3 shows the performance of Face Verification models for different face attributes such as Ethnicity, Attributes and Gender. Bias and fairness is measured by comparing the DoB_{fv} for models fine-tuned with CMU MultiPie and Synthetic faces. DoB_{fv} is greater for models trained with Synthentic faces. This is predominant at low FAR rates. This behaviour is not observed at high FAR rates. Our observations from the analysis of the results are as follows : (Observation-1 is drawn from experiment-1 and observations-2,3,4 were drawn from experiment-2)

- **Observation-1:** GANs are biased towards age group "20-29" and "White" faces.
- **Observation-2:** Face Verification models trained or fine-tuned with Synthetic faces exhibit bias for "race" attribute. This is confirmed by high DoB_{fv} for Synthetic faces when compared to CMU

MultiPie.

– **Observation-3:** Face Verification models trained or fine-tuned with Synthetic faces doesn't exhibit any bias for "gender" attribute.

– **Observation-4:** At, high FAR rates we don't observe bias (low DoB_{fv}). We hypothesize that although biases are present these are masked by high false acceptances.

These results were presented in ECCV 2022 Responsible Computer Vision workshop.

	GAR(%)	
FAR(%)	CMU Multi-PIE	Synthetic Faces
0.01	21.59	22.77
0.1	38.45	39.51
1	62.61	63.07
10	88.02	88.05

Table 3.2.1: GAR@FAR

	GAR(%)					
FAR(%)	CMU Multi-PIE			Synthetic Faces		
	Male(M)	Female(F)	Std	Male(M)	Female(F)	Std
0.01	22.77	19.98	1.97	22.71	23.04	0.23
0.1	41.47	36.6	3.44	40.62	38.36	1.60
1	66.23	60.55	4.02	64.63	61.43	2.26
10	88.85	87.54	0.93	88.54	87.53	0.71

Table 3.2.2: GAR@FAR for gender attribute

	GAR(%)									
FAR(%)	CMU Multi-PIE					Synthetic Faces				
	Asian(A)	Black(B)	Indian(I)	White(W)	Std	Asian(A)	Black(B)	Indian(I)	White(W)	Std
0.01	16.2	22.24	24.37	31.19	7.04	18.0	18.53	27.5	36.69	8.82
0.1	30.05	38.27	43.57	53.23	9.72	31.35	36.75	44.48	56.11	10.74
1	52.46	62.25	65.52	76.08	9.74	56.55	60.08	64.95	76.10	8.51
10	82.38	88.09	87.85	93.4	4.50	84.5	86.86	88.06	92.56	3.38

Table 3.2.3: GAR@FAR for ethnicity attribute

3.3 Results for fair distillation of generative models:

Around 5000 face images were generated by the distilled student generator. These were shown in 3.3.1. The attributes for all these images were obtained using pre-trained fairface classifier and plotted 3.3.4

Teacher GAN - Student GAN Faces

Figure 3.3.1: Teacher Student Faces

From figure 3.3.2 and 3.3.4, it is evident that imbalance improved after fair distillation. This can be specifically observed with race/race4 attribute. The proportion of faces for races other than white increased after fair distillation when compared to white faces.The imbalance improvement for age attribute is not substantial. The proportion of faces for different age groups almost remained the same even after fair distillation.

The quality of faces generated by the student GAN are not similar to the quality of faces from teacher GAN. However, the objective of the distillation process is fairness. However, further research is needed to improve imbalance for other attributes along with generated image quality.

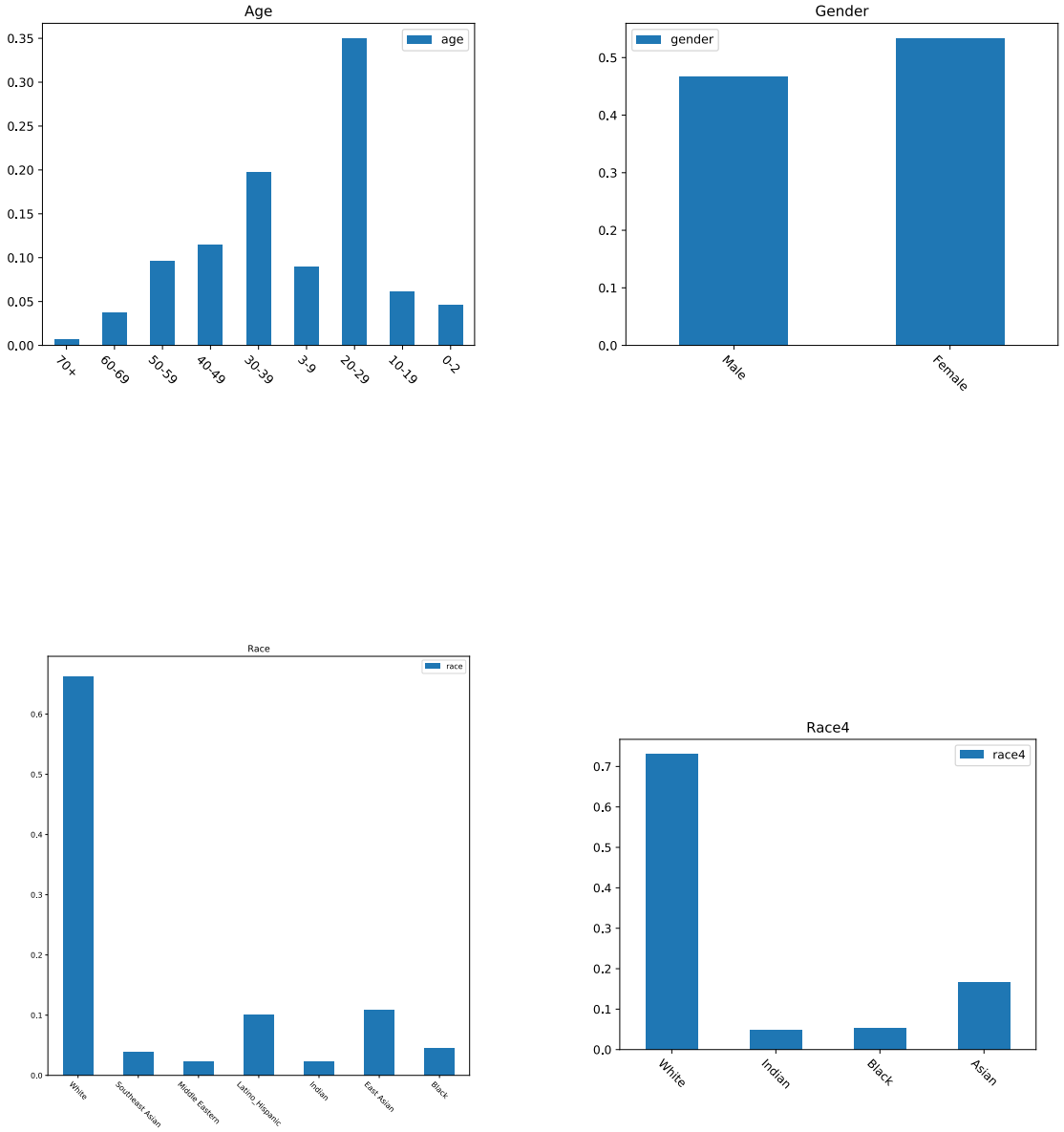


Figure 3.3.2: Proportion of faces in each sub-group for Age,Gender,Race and Race4 attribute for GAN generated synthetic faces(x-axis sub-groups and y-axis proportion)

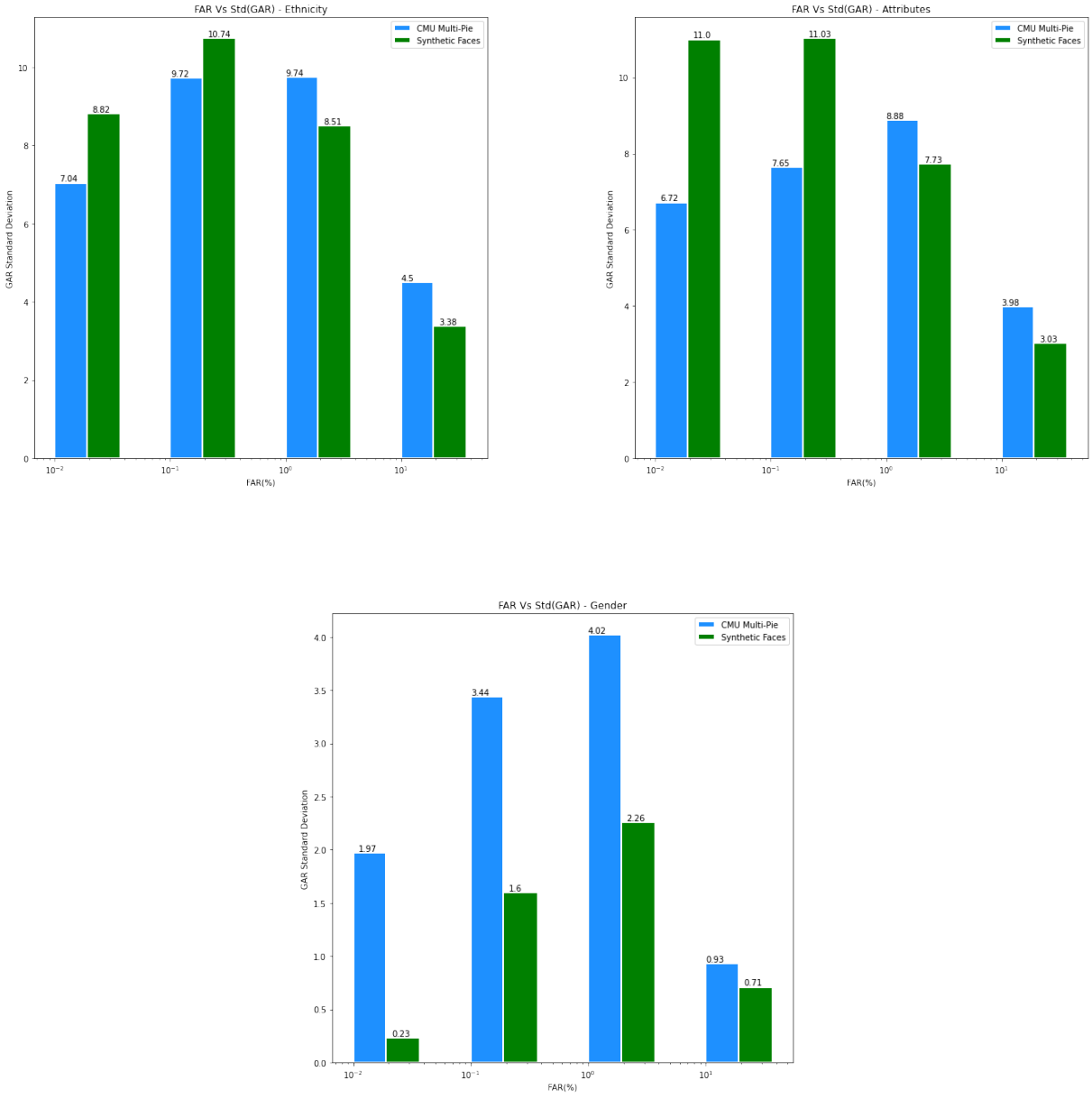


Figure 3.3.3: DoB_{fv} i.e Std(GAR @ FAR) for Ethnicity, Gender and Attributes with CMU Multi-Pie and Synthetic faces (smaller is better for bias)

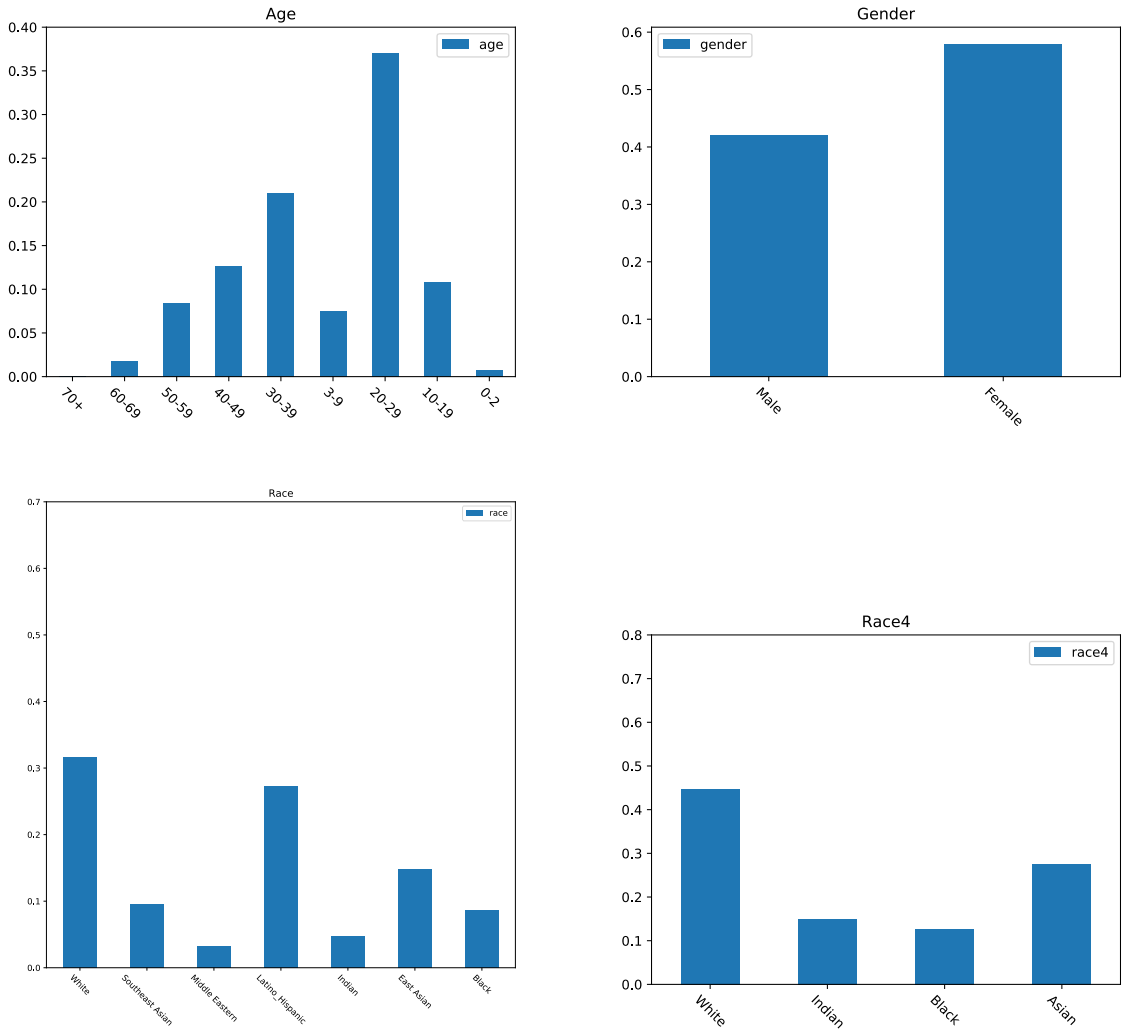


Figure 3.3.4: Proportion of faces in each sub-group for Age,Gender,Race and Race4 attribute for Fair Distilled GAN generated faces(x-axis sub-groups and y-axis proportion)

Chapter 4

Conclusion and Future Work

Recent advances in deep learning and computer vision showed impressive performance for Image recognition, specifically face recognition. However, these methods need high resolution images. In this work we demonstrated experimentally that performance of existing image recognition systems reduced drastically with low resolution or very low resolution images along with comparing the performance with high resolution images. Also existing state of the art techniques and algorithms for low(very) resolution image recognition were reviewed in the literature survey. We discerned from this that generative models are important components in low(very) resolution image recognition architectures.

On analysing GANs we demonstrated that generative models trained on ffhq dataset exhibit high bias for ethnicity sub groups. On the contrary we had shown that they doesn't exhibit bias for different gender sub groups. Along with this we also showed how biased generative models can cause disparate impact to downstream systems such as face verification systems. It is also evident from this that low(very) resolution image recognition architectures also need debiased generative models for fair recognition performance across different sub-groups.

To mitigate bias fair distillation approach is proposed. It was shown with this approach that fairness constraints at the data sampling level can help mitigate bias in student models. Additional research is also needed to establish the effectiveness of Shannon Diversity Loss as fairness contraint. In future, more research is needed particularly in the area of distillation of generative models for obtaining better image quality with fair student models. This helps obtain debiased compress GANs with can be deployed at edge across different sub-groups. We hope our work spurs further research.

Publications

[44] On Biased Behavior of GANs for Face Verification - Accepted at Responsible Computer Vision workshop, ECCV 2022.

References

[1] M. Singh, S. Nagpal, R. Singh, and M. Vatsa, "Derivenet for (very) low resolution image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[2] ——, "Dual directed capsule network for very low resolution image recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[3] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Proceedings. 1991 IEEE computer society conference on computer vision and pattern recognition.* IEEE Computer Society, 1991, pp. 586–587.

[4] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[5] X. Wu, R. He, Z. Sun, and T. Tan, "A light cnn for deep face representation with noisy labels," 2018.

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," 2018.

[7] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," 2018.

[8] F. Wang, J. Cheng, W. Liu, and H. Liu, "Additive margin softmax for face verification," *IEEE Signal Processing Letters*, vol. 25, no. 7, pp. 926–930, 2018.

[9] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "Cosface: Large margin cosine loss for deep face recognition," 2018.

[10] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," 2019.

[11] Y. Huang, Y. Wang, Y. Tai, X. Liu, P. Shen, S. Li, J. Li, and F. Huang, "Curricularface: Adaptive curriculum learning loss for deep face recognition," 2020.

[12] H. Phan and A. Nguyen, "Deepface-emd: Re-ranking using patch-wise earth mover's distance improves out-of-distribution face identification," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 259–20 269.

[13] K. Grm, W. J. Scheirer, and V. Štruc, "Face hallucination using cascaded super-resolution and identity priors," *IEEE Transactions on Image Processing*, vol. 29, pp. 2150–2165, 2020.

[14] C.-C. Hsu, C.-W. Lin, W.-T. Su, and G. Cheung, "Sigan: Siamese generative adversarial network for identity-preserving face hallucination," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 6225–6236, 2019.

[15] S. Ghosh, M. Vatsa, and R. Singh, "Suprear-net: Supervised resolution enhancement and recognition network," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2022.

[16] Q. Yun, Z. Cao, and E. Zhou, "Face hallucination using convolutional neural networks," Aug 2016.

[17] M. Singh, S. Nagpal, M. Vatsa, and R. Singh, "Enhancing fine-grained classification for low resolution images," *CoRR*, vol. abs/2105.00241, 2021. [Online]. Available: https://arxiv.org/abs/2105.00241

[18] M. Wang, R. Liu, N. Hajime, A. Narishige, H. Uchida, and T. Matsunami, "Improved knowledge distillation for training fast low resolution face recognition model," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 2019, pp. 2655–2661.

[19] O. A. Aghdam, B. Bozorgtabar, H. K. Ekenel, and J.-P. Thiran, "Exploring factors for improving low resolution face recognition," 2019.

[20] M. Liang, F. Yan, and Z. Shaowen, "Human facial feature detection method under low resolution."

[21] S. Nagpal, M. Singh, R. Singh, and M. Vatsa, "Deep learning for face recognition: Pride or prejudiced?" *arXiv preprint arXiv:1904.01219*, 2019.

[22] J. P. Robinson, G. Livitz, Y. Henon, C. Qin, Y. Fu, and S. Timoner, "Face recognition: too bias, or not too bias?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 0–1.

[23] X. Xu, Y. Huang, P. Shen, S. Li, J. Li, F. Huang, Y. Li, and Z. Cui, "Consistent instance false positive improves fairness in face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 578–586.

[24] V. V. Ramaswamy, S. S. Kim, and O. Russakovsky, "Fair attribute classification through latent space de-biasing," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9301–9310.

[25] S. Gong, X. Liu, and A. K. Jain, "Jointly de-biasing face recognition and demographic attribute estimation," in *European conference on computer vision*. Springer, 2020, pp. 330–347.

[26] C. Karakas, A. Dirik, E. Yalcinkaya, and P. Yanardag, "Fairstyle: Debiasing stylegan2 with style channel manipulations," *arXiv preprint arXiv:2202.06240*, 2022.

[27] N. Jain, A. Olmo, S. Sengupta, L. Manikonda, and S. Kambhampati, "Imperfect imaganation: Implications of gans exacerbating biases on facial data augmentation and snapchat selfie lenses," *arXiv preprint arXiv:2001.09528*, 2020.

[28] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network (2015)," *arXiv preprint arXiv:1503.02531*, vol. 2, 2015.

[29] A. Aguinaldo, P.-Y. Chiang, A. Gain, A. Patil, K. Pearson, and S. Feizi, "Compressing gans using knowledge distillation," *arXiv preprint arXiv:1902.00159*, 2019.

[30] Y. Ren, J. Wu, X. Xiao, and J. Yang, "Online multi-granularity distillation for gan compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6793–6803.

[31] T.-Y. Chang and C.-J. Lu, "Tinygan: Distilling biggan for conditional image generation," in *Proceedings of the Asian Conference on Computer Vision*, 2020.

[32] M. Du, S. Mukherjee, G. Wang, R. Tang, A. Awadallah, and X. Hu, "Fairness via representation neutralization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 091–12 103, 2021.

[33] T. Adel, I. Valera, Z. Ghahramani, and A. Weller, "One-network adversarial fairness," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 2412–2420.

[34] T. Jang, F. Zheng, and X. Wang, "Constructing a fair classifier with generated fair data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, 2021, pp. 7908–7916.

[35] L. E. Celis, A. Mehrotra, and N. Vishnoi, "Fair classification with adversarial perturbations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8158–8171, 2021.

[36] G. B. Huang, M. Mattar, H. Lee, and E. Learned-Miller, "Learning to align from scratch," in *NIPS*, 2012.

[37] Z. Cheng, X. Zhu, and S. Gong, "Surveillance face recognition challenge," *arXiv preprint arXiv:1804.09691*, 2018.

[38] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, pp. 539–546 vol. 1.

[39] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and vision computing*, vol. 28, no. 5, pp. 807–813, 2010.

[40] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[41] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Proc. NeurIPS*, 2020.

[42] K. Karkkainen and J. Joo, "Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1548–1558.

[43] Y. Deng, J. Yang, D. Chen, F. Wen, and X. Tong, "Disentangled and controllable face image generation via 3d imitative-contrastive learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[44] S. Kotti, M. Vatsa, and R. Singh, "On biased behavior of gans for face verification," *arXiv preprint arXiv:2208.13061*, 2022.