UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE ED
ELETTRICA E MATEMATICA APPLICATA

Corso di Laurea Magistrale in Ingegneria Informatica

Effect of Blind Face Restoration on Soft Biometrics Recognition

Relatori:
Prof. Antonio Greco
Prof. Mario Vento

Candidato:
Alfonso Cavallo
Mat. 0622701441

ANNO ACCADEMICO 2021/2022

Dedicato a mia nonna **Pasqualina**
nella speranza di essere sempre altezza
dell'amore che brillava nei suoi occhi
e dell'orgoglio che riempiva i suoi sorrisi.

CONTENTS

CHAPTER 1

INTRODUCTION

Artificial intelligence is becoming increasingly important in today's society. Thanks to its enormous potential we can solve more and more complex problems. According to Marco Somalvico, artificial intelligence studies the theoretical fundamentals, methodologies, and techniques to design hardware and software systems able to achieve a performance that a common observer would attribute exclusively to human intelligence. Through data and experience, those solutions can learn how to solve difficult tasks such as identifying a person or even creating music compositions and art.

Considering this, there are intelligent systems able to process data to make them an input more suitable for analysis or further processing steps. Popular examples are the Convolutional Neural Networks (CNNs) that can extract meaningful features from a digital representation of a face. In combination with the Generative Adversarial Networks' (GANs) capability of creating realistic fakes of face images, many attempts were made to create proficient Blind Face Restoration networks. Such artificial intelligence solutions are designed to remove real unknown corruption from faces and create a high-quality picture from a degraded version of the same.

Hopefully, a Blind Face Restoration network should be capable of

preserving individual soft biometrics while enhancing the data, in order to achieve a fine trade-off between a pleasurable result in terms of image quality and facial characteristics conservation. Such a compromise, if achieved, could represent a good solution to ease face recognition tasks like age estimation and classification of gender, ethnicity, facial emotion, and identity.

This thesis aims to design and implement an experimental framework to test whether Blind Face Restoration improves performance on soft biometrics recognition tasks. The results, collected on different datasets gathered in the wild, must be analyzed and compared to find the most determining factors in such neural networks' training.

1.1 Problem formulation

Blind Face Restoration aims to recover a high-quality face image from a face image corrupted by an unknown degradation. The utilization of this processing in artificial vision ranges from artistic purposes, such as the restoration of old multimedia material, to real-time security video surveillance and medical systems. Yet, the hardness of the task is prominent due to the irreversibility of most image corruptions, leading to ill-posed estimations of the inverse degradation process.

A formalization of the image degradation process is the following:

$$I_c = \phi(I, \delta) \tag{1.1}$$

where ϕ is the degradation model and δ represents its parameters. The objective of Blind Face Restoration is to find f, the inverse function of ϕ, such that:

$$f(I_c, \alpha) = \phi^{-1}(I_c, \alpha) = \hat{I} = I \tag{1.2}$$

where α represents the parameters of f. However, in real conditions, the degradation function ϕ is unknown, along with its parameters δ, making the task of finding ϕ^{-1} and α a challenging one. With this in mind, $\hat{I}$ is always an estimation of the original image I. The real approach, given the

reconstruction model f, aims to minimize a loss function $\mathcal{L}$ and find the best possible estimation by selecting α such that:

$$\alpha = argmin_{\alpha}\mathcal{L}(I, \hat{I}) \tag{1.3}$$

In literature, such an unknown degradation is simulated with mathematical models, that make researchers able to produce pairs I and I_c, useful to train restoration networks with supervised learning. The problem of low-quality pictures is analyzed deeply, with a specific focus on each possible corruption of the image. However, there are also generic degradation models, combining many different corruptions in a single transformation, and they are proven to provide a result that, despite being synthetic, is close to real-world degradation. The most accounted corruptions are Gaussian noise, JPEG compression, downsampling, and different kinds of blur.

1.2 Problem relevance in the information engineering context

Blind Face Restoration networks can help give new life to low-quality content and information, that could have been inevitably degraded by past times' technological limitations. Also, noise removal is a hot topic when it comes to in-the-wild images, that are an everyday problem in real-time video surveillance and medical monitoring. Modern engineering is always trying to find feasible solutions that grant a good compromise between speed and quality. Blind Face Restoration may be a good solution to improve performance on soft biometrics recognition tasks that require elaboration on real-world pictures. Some examples are facial emotion recognition to keep track of customer satisfaction levels or even people identification for security applications. However, a positive effect of restoration on those tasks is not obvious and no state-of-the-art study performed experiments to test such a correlation.

CHAPTER 2

STATE OF THE ART

State of the art is improving knowledge of Blind Face Restoration, with many different innovative network architectures and methods continuously being tested and released. Many different approaches are studied, each one analyzing different facial features, to grant the best restoration possible. Also, much of the researchers' attention is focused on image quality metrics, which provide a quantifiable measurement of the restoration quality. Some of them require a high-quality reference to retrieve a quality index, comparing the restored picture with the original one, while others give a score on real low-quality restored samples without the need for any reference. Furthermore, in the last years, extensive benchmarks and studies proposed new solutions to soft biometrics recognition tasks with always increasing performance in terms of error reduction and robustness.

2.1 Blind Face Restoration

It is possible to define Blind Face Restoration through the following statement:

"Blind face restoration aims at recovering high-quality faces from the low-quality counterparts suffering from unknown degradation, such as low-resolution, noise, blur, compression artifacts, etc."

Wang, Xintao, et al. [1]

According to this definition, Blind Face Restoration must be able to clean a face picture without knowing the degradation and its parameters. To better achieve this objective, many studies showed that adding significant additional priors may enhance models' performance. The methods based on geometric priors add heatmaps, landmarks, and many other components to the equation. PSFRGAN [2], for example, progressively enhances face quality by being constantly aware of a face semantic mask. Reference-based restoration methods, instead, rely on a high-quality face reference of the subject portrayed by the degraded input image. ASFFNet [3] extracts an attention map from a guidance image and a landmark map to better optimize the restoration. GFRNet [4], instead, uses a warping network to make the reference face match spatially the degraded one, before passing both of them to a decoder. Those methods achieve good performance but a reference face is not always accessible. Some other solutions rely on high-quality dictionaries to enhance their performance. DFDNet [5] uses RoIAlign pooling to extract face components from different scales VGG features, and K-means for clustering, to build dictionaries of components for helping the restoration process. Differently, in RestoreFormer [6], the dictionary is reconstruction-oriented and learned with vector quantization by a high-quality face generation network. Networks based on generative priors try to optimize the extracted latent vector and improve the generation process of the restored face. GFPGAN [1] combines generative facial priors extracted by pre-trained face GANs with spatial features transform layers. GPEN [7] integrates a trained face GAN into a U-shaped encoder-decoder architecture as a decoder. Other novel solutions are the RMM framework [8] which combines the learnable universal priors and unsupervised wavelet memory, and FP-FRN [9] that's an adversarial network

based on multi-scale feature extraction.

2.2 Image Degradation

Image degradation is a wide field studying all kinds of corruption that can occur during image acquisition and processing. It has been deeply studied as well as restoration tasks required to eliminate the effects of certain corruptions, individually. A description of the most studied reconstruction tasks in literature is provided by the following list:

- **Image Denoising** is about the removal of many different noises from an image. Noise is a statistical distortion that usually interferes during the data acquisition step. It usually affects sensors with small random variations or spontaneous peaks due to environmental causes like temperature, luminance, or inner unreliability of the sensors themselves. Classic artificial vision solutions require a lot of hyperparameters tuning to work and they're easily surpassed by deep neural networks performance. While some of those models try to estimate the noise to remove it, many other approaches are based on a blind restoration of the sample.

  The most popular noise degradation model is the Gaussian noise which can be modeled as follows:

  $$I_d = I + n_\sigma \tag{2.1}$$

  where n_σ is the random noise generated by the Gaussian distribution $\mathcal{N}(0, \sigma^2)$.

- **Image Deblurring** is about the removal of blur from an image. Some of the reasons behind this distortion are defocusing and optical aberrations, but, more often, it's caused by the movements of the camera or subjects in the field of view. Image deblurring task follows almost the same rule of image denoising, so the most popular solutions are non-blind models,

with prior knowledge about the blur hyperparameters and parameters, and the even more popular blind models that will try to estimate them.

From a mathematical point of view, blur can be modeled as a convolution between the image I and a blur kernel K, which may vary depending on the type of blur.

$$I_d = I * K \tag{2.2}$$

- **JPEG Deblocking** is about the removal of JPEG artifacts from a compressed image. Given an image, it is possible to compress its information content at different levels of quality for an economic storage of the same. Depending on such level, the compression may be more or less lossy, leading to the generation of block artifacts that makes, for example, high-frequency contents look blurry. Deep neural networks are usually employed for deblocking tasks to infer lost information.

- **Super-resolution** is about the enhancement of a low-quality image resolution to a higher level of detail. The simplest way to formalize information loss is the downsampling model, usually represented as a bicubic interpolation with a variable scale factor s:

$$I_d = (I) \downarrow_s \tag{2.3}$$

  However, in the real case scenario, super-resolution is not that simple and has to eliminate also the contributions of many different corruptions, as well as bring data to a higher resolution.

Blind Face Restoration aims to give the better reconstruction possible given an unknown degradation, specifically on face images. For this reason, models trained to achieve this task have to be trained on all the degradations previously described. To achieve this objective, those models are usually trained on high-quality datasets that are synthetically degraded by applying all the corruptions, usually combined in a single general degradation model. The most famous general degradation formalization in literature was proposed

by *Li et al.* [4] and it's thought to merge all the disturbances coming into play in a long-distance acquisition. Its mathematical formulation is the following:

$$I_d = ((I * K_\delta) \downarrow_s + n_\sigma)_{JPEG_q} \tag{2.4}$$

where K_δ is the isotropic Gaussian blur with standard deviation δ representing the defocus effect, n_σ is the Gaussian noise with standard deviation σ added lastly to simulate AWGN, s is the scale factor of the bicubic downsampler, and q is the quality of JPEG compression employed for economic storage. Although many variations of this formalization have been tested in the state of the art, with different parameters, noise, and blur kernel, this model proved to be a solid estimation for real unknown degradation.

2.3 Image Quality Metrics

The problem of an image's quality measure can be approached in many different ways, depending on the availability of a high-quality reference. Such a condition leads to a division in image quality metrics into two macro groups:

"Specifically, full-reference image quality assessment (FR-IQA) algorithms possess full information for both the distorted image and the reference image, while no-reference image quality assessment (NR-IQA) methods predict perceptual quality exclusively based on the distorted image."

Domonkos Varga [10]

2.3.1 Full-Reference Quality Metrics

Full-Reference quality metrics measure the quality of a distorted image by the similarity with a high-quality reference of the same.

- **Mean Square Error** ($\mathcal{L}_2$) [11] is the simplest quality metric. It measures the average square error on all the color channels of the picture.

Its mathematical formula is the following:

$$MSE(I_d, I) = \frac{1}{HWC}\|I - I_d\|^2 \tag{2.5}$$

where W and H are the shape of the image, C is the number of color channels, I is the high-quality reference, and I_d is the degraded version of the same.

- **Mean Absolute Error** ($\mathcal{L}_1$) [11], following the same principle of MSE, measures the average absolute error on all the color channels of the pictures:

$$MAE(I_d, I) = \frac{1}{HWC}\sum_i |I_i - I_{d_i}| \tag{2.6}$$

- **Peak Signal-to-Noise Ratio** [11] is as simple as the previous described metrics. It evaluates the ratio between max pixel intensity and the distortion power measured with MSE:

$$PSNR(I_d, I) = 10\log_{10}(\frac{max(I)}{\sqrt{MSE(I_d, I)}}) \tag{2.7}$$

This metric, as for MSE and MAE, measures the error on the different color channels pixel by pixel. All of them do an excellent job of providing a good measure of the raw difference between the reference and the degraded image but fail to give an evaluation of quality that's near to the human perception.

- **Structural Similarity Index** [11] is a more complex quality metric that adds many variables into the equation.

$$SSIM(I_d, I) = l(I_d, I)^\alpha + c(I_d, I)^\beta + s(I_d, I)^\gamma \tag{2.8}$$

where l, c, and s measure the comparison, respectively, in terms of luminance, contrast and structure, while α, β and γ are exponential weights. Its result is a number between 0 and 1, where 0 means no similarity at all while 1 means the two pictures are the same. This metric is perception-based so it takes into account the dependency of pixels in the image, especially when they are very close. For its ability

to analyze the structural information of the picture, this score is nearer to the way humans perceive quality in an image.

- **Learned Perceptual Image Patch Similarity** (LPIPS) [12] is a measure based on a deep visual representation of the pictures. Some studies proved that Convolutional Neural Networks extracted features, largely used in image generation as perceptual losses, provide also excellent scores to evaluate image quality.

2.3.2 No-Reference Quality Metrics

No-Reference quality metrics measure the quality of a distorted image without the need for a high-quality reference.

- **Mean Opinion Score** (MOS) is the simplest quality measure. It is totally opinion-aware because it is the average on the direct human evaluations of the quality, formalized as votes ranging from 0 to 5.

- **Fréchet Inception Distance** [13] measures the quality of an image by comparing the distribution caught inside it with the distribution of a set of real images. Its formalization is expressed as follows:

$$FID(I_d) = \|\mu_{I_d} - \mu_D\|_2^2 + tr(\sigma_{I_d}^2 + \sigma_D^2 - 2\sqrt{(\sigma_{I_d}^2 \sigma_D^2)}) \qquad (2.9)$$

  where the distribution are $f(I_d) = \mathcal{N}(\mu_{I_d}, \sigma_{I_d}^2)$ and $f(D) = \mathcal{N}(\mu_D^2, \sigma_D^2)$, D is the reference dataset, $\mathcal{N}$ is the distribution of the image, and tr is the trace. Differently from Full-Reference quality metrics, the ground-truth dataset makes it possible to get a quality score even without the same image's high-quality version. However, according to Chong, Min Jin et al. [14] study, this metric appears to be a biased and unreliable score.

- **Blind/Referenceless Image Spatial Quality Evaluator** (BRISQUE) [15] gives a score based on a Support Vector Regressor trained on an image database labeled with DMOS values. It is opinion-aware because it is built from human evaluations.

- **Naturalness Image Quality Evaluator** (NIQE) [16] evaluates the quality score as the deviation of the image from the statistical regularities extracted by a reference dataset of natural images, without, exposition to distortion, indeed. For this reason, NIQE is completely blind and opinion-unaware.

- **Perception-based Image Quality Evaluator** (PIQE) [17] calculates the no-reference quality score for an image through block-wise distortion estimation. It is an unsupervised opinion-unaware method.

2.4 Soft Biometrics Recognition

Soft biometrics recognition is one of the most popular artificial vision challenges of the last years. Among the huge set of possible application fields, there are security surveillance, medical, commercial, administrative, and many others. For a better understanding of the relevance of the problem, the following definition briefly describes what a soft biometric is:

"Soft biometric traits are physical, behavioral or adhered human characteristics, classifiable in pre–defined human compliant categories. These categories are, unlike in the classical biometric case, established and time–proven by humans with the aim of differentiating individuals. In other words the soft biometric traits instances are created in a natural way, used by humans to distinguish their peers."

Dantcheva et al. [18]

With this in mind, soft biometrics are individual traits that can be classified and used to distinguish humans. Specifically, face soft biometrics are those individual traits that can be identifiable by analyzing an individual face, such as gender, age, facial emotion, ethnicity, and identity. As for many others "human tasks", deep neural networks prove to be the best models for

such problems. Convolutional Neural Networks (CNNs), in particular, are the most effective solutions for those tasks, thanks to their capability of considering neighbor pixels' inter-dependency.

Despite many face datasets gathered in controlled environments being available, data collected and labeled in the wild are still very few, making the problem of soft biometrics recognition very challenging in real contexts. Furthermore, real environments are always exposed to many different sources of corruption due to technology limitations and real-world phenomena. With degradation being a key factor in soft biometrics analysis, the robustness of the model becomes a central objective of the training phase.

Gender recognition in the wild is explored in [19], considering synthetic corruption and in-the-wild faces, introducing also MiviaGender, a dataset of faces gathered in a real environment. Greco et al. [20] also provide an extensive benchmark of corruption effects on a facial emotions dataset collected in real scenarios. Age estimation is approached as a regression problem in [21]. Here, knowledge distillation is proposed to overcome the lack of annotations by transferring the performance from more complex teacher models to simpler students. Such models, as for the previously described ones, have also been tested on degraded and in-the-wild samples. In [22], a new dataset VMER is proposed to overcome the lack of faces data labeled with ethnicity information, along with an extensive cross-evaluation of the state-of-the-art networks on different datasets. Also, a multi-tasking approach has been used to estimate all the soft biometrics in [23].

The identification problem, differently from the previous tasks, focuses on the extraction of meaningful and distinctive features. Such features can be compared with a similarity measure to asses, with a certain confidence, whether two faces represent the same individual or not. VGGFace descriptor [24] is one of the most popular descriptors available in the state of the art.

2.5 Progression to the state-of-the-art

Blind Face Restoration and, in general, face degradation removal are very active branches of research, due to their wide range of possible utilization. Many network architectures have been deeply analyzed to find the best possible solution to the problem. In the same way, soft biometrics recognition has been deeply studied and innovative solutions have been tested in various challenging conditions. However, despite both those fields being very active, to the best of my knowledge, there is no study trying to analyze the impact of Blind Face Restoration on soft biometrics recognition tasks. This thesis aims to provide a quantitative and qualitative analysis of this effect, by comparing different architectures on different datasets. Furthermore, different hypotheses are tested to find an explanation for the resulting behavior.

CHAPTER 3

PROPOSED METHOD

The main objective of this thesis is to build an experimental framework able to compare Blind Face Restoration networks' performance, which is measured on three key points:

- Image Quality of the restored face.
- Contribution to the performance of soft biometrics recognition networks when tested on restored datasets.
- Elaboration speed calculated on restoration runs.

The design of the experiment is conducted downstream of the state of the art analysis explored in Chapter 2. The proposed method explores all those key factors on (i) different Blind Face Restoration networks, (ii) soft biometrics recognition networks, and (iii) datasets gathered in different conditions. Finally, the gathered results are analyzed to find convincing motivations behind the resulting behavior. Such workflow is divided into two main macro activities aiming to gather two sets of information: *Blind Face Restoration quality* and *Restoration Effects on soft biometrics recognition.* Figure 3.1 and Figure 3.2 provide a graphic representation of the two main workflows designed to carry on such an experiment.

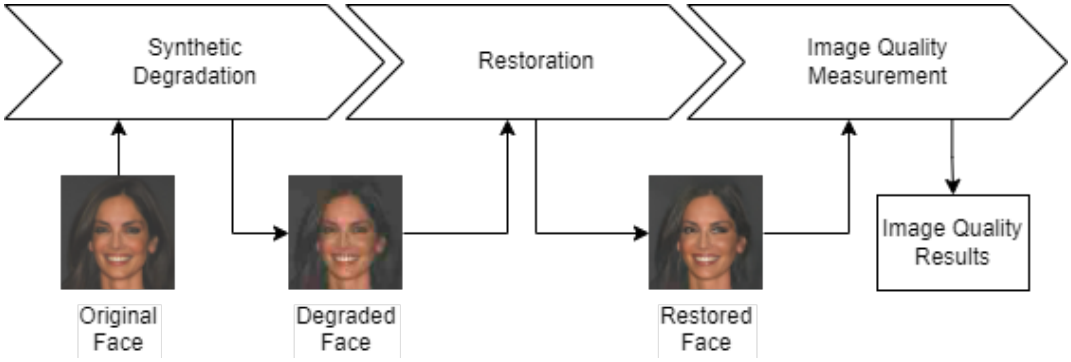

Figure 3.1: Image Quality experimental framework.

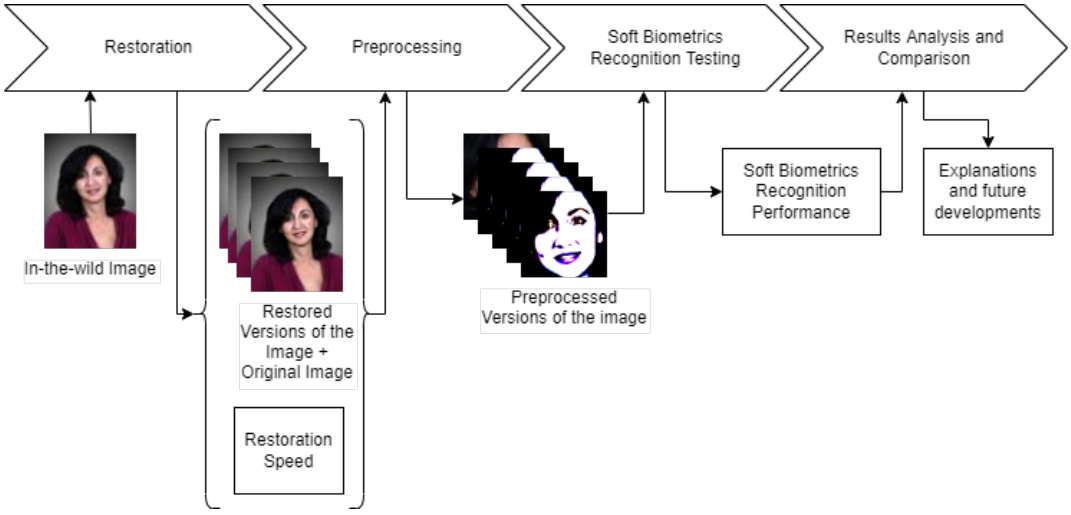

Figure 3.2: Experimental framework for Blind Face Restoration effect on soft biometrics recognition.

The complete framework is characterized by the following steps:

1. High-quality available data are synthetically degraded;
2. The restoration of synthetically degraded faces is run for each of the Blind Face Restoration networks integrated into the framework;
3. Quality is measured by comparing degraded and restored samples through full-reference quality metrics;
4. The restoration of soft-biometrics labeled datasets is run for each of the Blind Face Restoration networks integrated into the framework;

5. Many pre-processed versions of the restored and the original data are generated to match each soft biometrics recognition model' specification. Also, restoration speed is calculated;

6. A measurement of recognition performance is retrieved on both pre-processed original and restored data;

7. The results are analyzed, compared and discussed to find out the reasons behind them and possible future innovations.

3.1 Dataset

The dataset employed in this experiment is a collection of seven different datasets among the most popular in the state of the art. Each of them provides information about a subset of the target soft biometrics for the experiment, which are: gender, age, ethnicity, facial emotion, and identity. However, only CelebA-HQ includes high-quality faces employable for restoration quality analysis. A brief description of each one is presented in the following list:

- **CelebA-HQ Test** [25] is composed of 30,000 high-quality faces samples with a shape of 1080x1080. It is one of the most popular test sets for Blind Face Restoration quality analysis.

- **VggFace2 Test** [26] is a large dataset of 169,371 faces and 500 identities. A subset composed of the first 207 identities in alphabetic order by name, with a total dataset size of 70,909 samples. It is enriched with VMER [22] and VMAGE [21] labels and provides information about the gender, age, ethnicity, and identity of each face. It is the largest and the most complete dataset for available labels.

- **UTKFace** [27] contains 24,102 faces gathered in the wild labeled with gender, age, and ethnicity.

- **LFW** [28] contains 13,234 in-the-wild faces already cropped and aligned labeled with gender information.

- **MiviaGender Test** [19] contains 200 gray-scale faces gathered in a real environment with gender labels.
- **RAF-DB Test** [29] contains 2,869 faces gathered in the wild with a wide range of facial emotions, labeled and divided into seven classes.
- **FairFace Test** [30] contains 10,954 balanced on 7 different ethnicities which can be further reduced into 4 more general classes. Along with ethnicity, data are provided with gender labels.

Most of the datasets provide faces gathered in uncontrolled conditions, which are particularly interesting when it comes to analyzing the effect of the restoration in real application contexts.

3.2 The Blind Face Restoration networks

The whole experimental framework is tested on three network architectures that propose different approaches to the Blind Face Restoration problem. For a fair comparison, the test includes only models requiring nothing else than the low-quality image to perform its restoration. Models requiring guidance or additional inputs to perform the task are excluded from this experiment. For the same purpose, all the selected models have been trained on synthetically degraded versions of the same dataset, **FFHQ**. It is a high-quality face dataset extracted from the website *Flickr.com*. This subsection explores the chosen model for such an experiment.

3.2.1 PSFRGAN

PSFRGAN [2] is a "progressive semantic-aware style transformation framework". As visible in Figure 3.3, PSFRGAN is a pyramid of inputs at different scales that gradually upscale the input face to a higher resolution. After each de-convolution layer, the upscaled output is modulated with a style transformation that is always semantic-aware, having in input a semantic mask of the face. The framework also provides a pre-trained Face Parsing Network to

get semantic information from faces. PSFRGAN's training takes into account a Reconstruction Loss based on $\mathcal{L}_2$ loss and an Adversarial Loss. It also considers an original Semantic-Aware Style Loss which is a Gram Matrix Loss calculated on VGG19 features, for different layers, on each separated semantic part of the face.

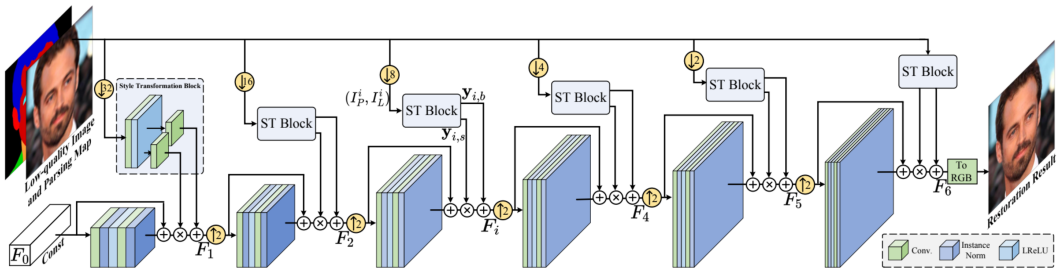

Figure 3.3: PSFRGAN architecture.

3.2.2 GFPGAN

GFPGAN [1] bases its face restoration power on Generative Facial Priors, after which the network is named. Figure 3.4 is the graphic representation of its architecture. The restoration process is structured as a pyramid of generative priors, bridged with optimized latent code, and fed to a Channel-Split Spatial Features Transformation block, which modulates the feature, given the spatial information. Latent features and spatial features are both extracted from a U-Net architecture, from the middle layer and the decoder de-convolutional layers respectively. GFPGAN's training considers a Reconstruction Loss based on $\mathcal{L}_1$ loss and an Adversarial Loss. It also includes a Component Loss, based on a Style Loss, and a Discriminator Loss, optimized on the left eye, right eye, and mouth. Finally, an Identity Loss is optimized as the distance with ArcFace features.

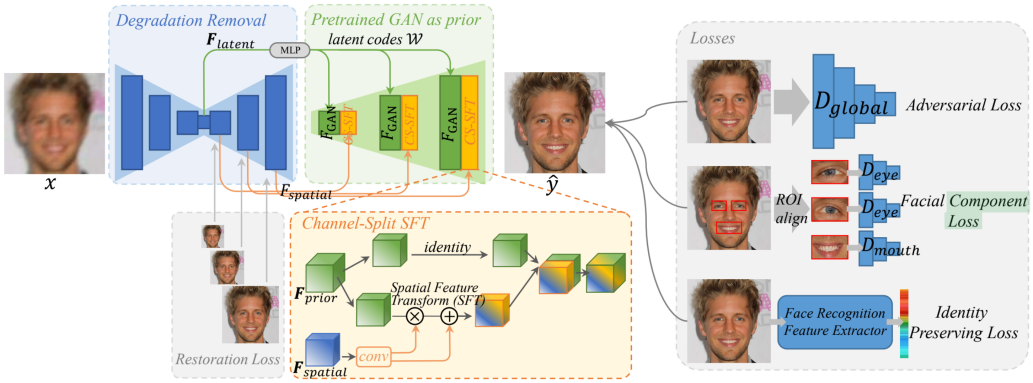

Figure 3.4: GFPGAN architecture.

3.2.3 RestoreFormer

RestoreFormer [6] face restoration relies on a high-quality features dictionary that is trained offline. Figure 3.5 illustrates how a first encoder block extracts latent features for the input face. Such features are compared with an HQ dictionary and the nearest prior is chosen. The high-quality face prior selected this way is passed to a series of two Multi-Head Cross-Attention blocks, which are also aware of initial extracted features. In the end, the output of the last block is upscaled by a decoder block to a high-quality representation. RestoreFormer's training considers a Reconstruction loss based on $\mathcal{L}_1$ loss and an Adversarial loss. It also includes an Identity Loss based on ArcFace. Also, it optimizes a Discriminator Loss and a Style Loss for the left eye, right eye, and mouth individually. Furthermore, a Reconstruction Loss based on $\mathcal{L}_1$ loss and an Adversarial Loss is also considered in the dictionary's training.

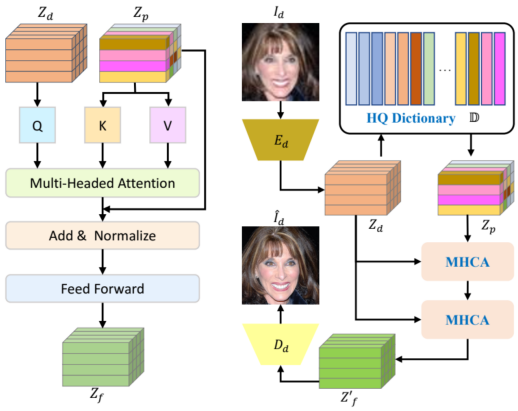

Figure 3.5: RestoreFormer architecture.

3.3 Synthetic Degradation

The high-quality data of CelebA-HQ need to be degraded to a lower quality before the restoration. The selected corruption framework is the general degradation model in Formula 2.4. Each image of the dataset is degraded with parameter values randomly sampled in the following intervals proposed by *Li et al.* [4]:

- **Blur Kernel**: $\rho \in \{0, 1 : 0.1 : 3\}$ where ρ is the std. deviation of K;
- **Downsampling**: $s \in \{1 : 0.1 : 8\}$ where s is the downscale factor;
- **Noise**: $\sigma \in \{0 : 1 : 7\}$ where the Gaussian noise is $\mathcal{N}(0, \sigma^2)$
- **JPEG compression**: $q \in \{0, 10 : 1 : 40\}$ where q is the quality with $q = 0$ is the losslessly compression

where interval is expressed as $begin : step : end$. An example of a degraded picture can be observed in Figure 3.6.

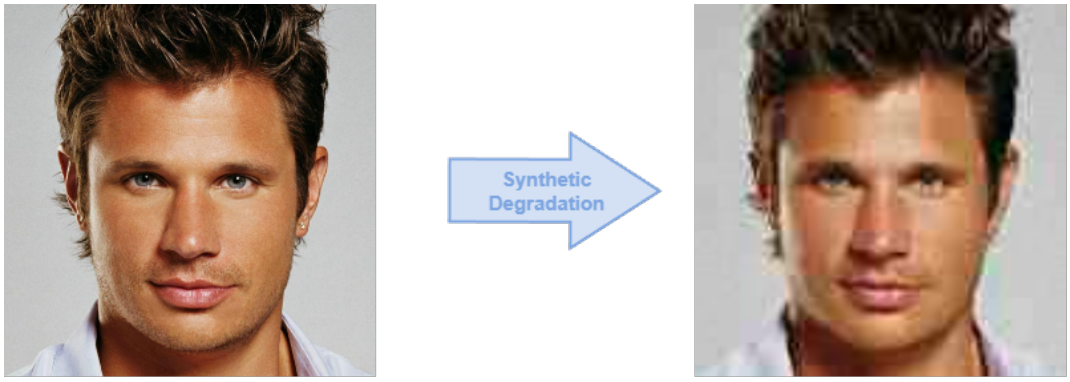

Figure 3.6: Face degraded with the general degradation model.

3.4 Soft biometrics recognition networks

In this section, I will provide a brief description of the target soft biometrics recognition tasks and the networks chosen to analyze the restoration effect on performance. The considered tasks are described in the following list:

- **Gender classification** is a binary classification problem based on two classes: Male and Female. The network employed to test the effect of restoration on such a task is SENet, the best network proposed by Greco et al. in [19].

- **Age estimation** is a regression problem with age included in [0, 120]. The network employed to test the effect of restoration on such a task is SENet, the best network proposed by Greco et al. in [21], trained on VMAGE through knowledge distillation. However, the output of this network is clipped in the interval [0, 100].

- **Ethnicity classification** is a multi-class classification problem based on four fundamental ethnicities: Caucasian Latin, African American, East Asian, and Asian Indian. The network employed to test the effect of restoration on such a task is VggFace, the best network proposed by Greco et al. in [22], trained on the VMER dataset.

- **Facial Emotion classification** is a multi-class classification problem

based on seven fundamental emotions: surprise, fear, disgust, happiness, sadness, anger, and neutral. The network employed to test the effect of restoration on such a task is the multitask Seresnet B, the best network proposed by Foggia et al. [23].

- **Identification** is a multi-class classification problem in which each identity represents a class. The identification problem is approached by splitting the test data into two equal-sized test set and support set, with random sampling. A constant seed is defined for the pseudo-random generation. The embeddings of each face in the test set are compared with the embeddings of each face in the support set. Finally, each face is classified as the same identity as the one in the support set it has the lowest cosine similarity 3.1 with. The network employed to extract the embeddings for each face is the notorious VggFace descriptor proposed by Parkhi et al. [24].

$$similarity(a, b) = \frac{a \cdot b}{\|a\|\|b\|} \tag{3.1}$$

3.5 Pre-processing

Different neural networks require specific data pre-processing to fully exploit their potential and fairly compare their performance. All the single-task networks employed to analyze gender, age, ethnicity, and identification share the same pre-processing. Differently, the multi-tasking network Seresnet B used for facial emotion classification requires another specific pre-processing. The complete pre-processing algorithm is reported in Figure 3.7.

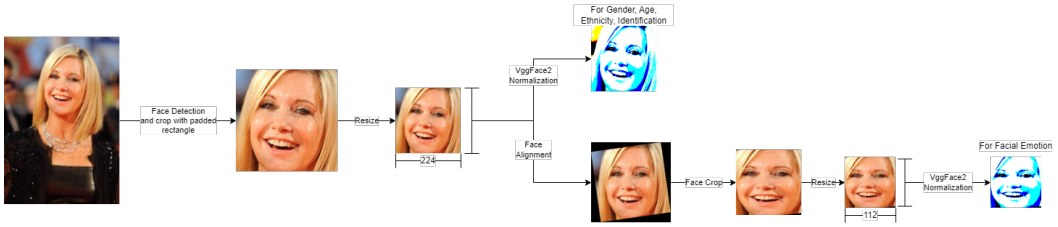

Figure 3.7: Image pre-processing algorithm.

For the process to be carried on correctly, the face must be loaded in RGB format. In order to optimize the framework, the first part of the pre-processing is shared among all the models:

1. *Faces cropping (with square padding)*

   The image is passed to an SSD [31] detector with ResNet-10 as a backbone network, which proved to be a very effective CNN for face detection. If more than one face box is found by the model, only the one with the highest confidence score is cropped. Lastly, the cropped box is padded to a square window;

2. *Resizing*

   The cropped square is resized to a size of 224×224 pixels;

As regards **multi-task Seresnet B**, the pre-processing algorithm requires a few more steps:

1. *Face alignment*

   The image is aligned thanks to the output of the *dlib* 64 face landmarks predictor;

2. *Face cropping (without padding)*

   The image is further cropped thanks to the SSD detector. This time, no padding is applied to obtain a square-shaped box;

3. *Resizing*

   The cropped face is resized to a size of 112×112 pixels.

However, all the networks require the following final step for the pre-processing:

1. *Normalization*

   VggFace2 pre-processing is applied to the face, normalizing face colors according to the mean value calculated on the VggFace2 dataset.

3.6 Tools and Technology

3.6.1 Hardware

The hardware resources available are located on the cloud, on a workstation reachable through the SSH protocol. The workstation provides 8 GB of RAM on a *Nvidia Quadro RTX 8000 GPU*. A big *cloud storage* and a *fast SSD* for fast data reading and writing during the elaborations. Employed CPU is *Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz*. Also, an *Nvidia TITAN Xp GPU* with 12 GB of RAM was available for a limited amount of time.

3.6.2 Software

The coding tools and development environments used are *Visual Studio Code* and *Pycharm*. The version control system adopted is *GitHub*. The package management system used to handle different python virtual environments is *Anaconda*. Filezilla and MobaXTerm have been used as SSH interfaces with the server. The most used software libraries are: *OpenCV*, *Tensorflow*, *Keras*, *Scikit-learn*, *PyTorch*, *pytorch_ssim*, *torchmetrics*, *CudaToolkit*, *Numpy*, *Cupy*, *keras-vggface*, *Pillow*, *Pandas*, and *h5py*. For documentation *Microsoft Word*, *Overleaf*, and *draw.io* have been used. *Microsoft Excel* and *Google Sheets* are the software used for spreadsheet management. *Microsoft Powerpoint* was used to produce presentation material. Finally, *Google Drive* was used to store the processed datasets.

CHAPTER 4

EXPERIMENTAL RESULTS

The main purpose of the proposed experimental framework is to evaluate the performance of the selected Blind Face Restoration networks. Such results have been gathered on networks' elaboration speeds, image reconstruction quality, and effects on soft biometrics recognition tasks.

Firstly, the metrics to analyze the quality of the restoration and the performance of soft biometrics recognition tasks are defined.

At the end of the restoration process, logs are analyzed to extract information about the average elaboration speeds of the models relative to the available hardware, on a large number of samples. Then, I provide a quantitative evaluation of the quality of the restoration process. Subsequently, I carry on a qualitative comparison. Lastly, after defining the soft biometrics recognition performance metrics, I gather the results on the original dataset and all the restored versions of the same. A comparison is performed along with an analysis to spot the causes that lead, eventually, to the resulting behavior.

4.1 Evaluation Metrics

In this section, I present the evaluation metrics selected to measure the performance of the restoration and the effect on soft biometrics recognition tasks.

4.1.1 Image Quality Metrics

The selected image quality metrics, given the availability of a high-quality reference dataset, are **MSE** and **PSNR**, which provide a great raw measure of the distance between the restored image and the original one, based on the mean square error calculated pixel by pixel. To also measure the human-perceived quality of the image with a more complex measure, **SSIM** is the employed metric. It provides an easy-to-read index, ranging 0 to 1.

4.1.2 Soft biometrics recognition metrics

The performance analysis of soft biometrics recognition tasks, for the sake of an easier comparison between the results, is carried on with **Accuracy Score** and **Balanced Accuracy Score** on all the classification tasks, so gender, ethnicity, facial emotion classification, and also identification, as it was implemented as a classification problem. Age estimation is a regression problem and the **Mean Absolute Error (MAE)** is the selected metric to measure its performance.

4.2 Restoration Elaboration Speed

The elaboration speed of the Blind Face Restoration networks is calculated on average considering the first 1,000 batches of size 20 of VggFace2 Test. The first batch is excluded and the speed is related to the available GPU *Nvidia Quadro RTX 8000 GPU* and the available CPU *Intel(R) Xeon(R) Silver 4210R CPU @ 2.40GHz*. Table 4.1 shows off the final results measured in frames per second (FPS).

Table 4.1: Blind Face Restoration networks' speed in FPS.

Network	Elaboration Speed ↑
GFPGAN	0.4987
PSFRGAN	0.3311
RestoreFormer	0.4569

GFPGAN is the faster network, followed by RestoreFormer only slightly slower. The slowest model considered is PSFRGAN. Such a result is motivated by the architecture of PSFRGAN which is very deep because of its progressive upscales. Differently from the others, PSFRGAN must also run a Semantic Segmentation on the face before restoring it, making the entire process very slow. GFPGAN being faster than RestoreFormer, instead, can be explained with the presence of the high-quality face priors dictionary that is considered in the second one, but not in the first one.

4.3 Restoration Quality Analysis

In this section, I analyze the quality of the restored face for each tested network. At first, I provide a quantitative analysis based on quantifiable metrics, then I provide a qualitative comparison of some examples.

4.3.1 Quantitative Analysis

The quantitative analysis results, as described in Section 4.1.1, are carried on with MSE, PSNR, and SSIM. Table 4.2 shows off the results for each restoration.

Table 4.2: Image quality of the restored faces.

Network	MSE ↓	PSNR ↑	SSIM ↑
GFPGAN	128.0634	28.8600	0.7848
PSFRGAN	208.0450	25.5536	0.7637
RestoreFormer	255.7733	25.0400	0.7334

The symbol ↑ means the result gets better when the value increases, while ↓ has the opposite meaning. Blue color marks the cell with the best value on a column while red color marks the worst. This standard is applied to all the tables in the thesis.

According to the results, GFPGAN is largely superior to the other blind face restoration networks, on all the quality metrics. RestoreFormer, instead, has the worst reconstruction performance. The approach based on generative priors leads to better results in terms of both frame rate and reconstruction quality.

4.3.2 Qualitative Analysis

Qualitative analysis is performed for a better understanding of the quantitative results. Figure 4.1 illustrates a comparison grid of the Blind Face Restoration, on some examples that diverge for gender, ethnicity, facial emotion, and head position.

Figure 4.1: Qualitative comparison of the Restoration.

The qualitative analysis confirms the results. GFPGAN restoration achieves smoother results, which are less corrupted and have a more pleasurable effect on human perception. PSFRGAN and RestoreFormer, instead, despite providing a qualitatively good result, aren't able to fully remove distortion and the result is still contaminated by some minor corruptions.

4.4 Soft Biometrics Recognition Performance

The main purpose of the experiment, as already mentioned, is to analyze whether Blind Face Restoration has an impact on recognition tasks, on target soft biometrics like gender, age, ethnicity, facial emotion, and identity. To prove this point, Accuracy and Balanced Accuracy scores are calculated on the original dataset and each one of the three restored versions of the same, for each dataset integrated into the framework. As it will be clear in the next subsection, **Blind Face Restoration doesn't preserve all soft biometrics**. In order to comprehend the motivations behind such behavior,

some hypotheses are initially presented and subsequently analyzed to prove their correlation with the results. The following list provides a description of such assumptions:

- Neural network performance is strictly related to the data. If the training is biased, a Blind Face Restoration model may be more likely to produce faces with biometrics similar to the most occurring soft biometrics in training. All three tested networks share a common training set, FFHQ. It is gathered on *Flickr.com* in controlled conditions and inevitably inherits the bias of the website itself. In order to check if the unbalance in data pours out also into the restored data, a study of the bias is conducted on the soft biometrics predictions obtained after the reconstruction.

- Neural networks' skill in specific tasks mainly depends on the loss functions considered during the training step. If the parameters of a model haven't been tuned to optimize a specific loss related to a certain task, the network will more likely ignore it and focus on the objective functions that have, instead, been defined. This means that, if no soft-biometrics-preserving loss is considered during the training of the network, the model may prefer to drop some soft biometrics information to achieve a better image quality, or simply just be unable to preserve it.

4.4.1 Gender Classification

The results of the gender classification task are retrieved with gender SENet. For a better comprehension of the data, the gathered results are graphically represented as line plots. Figure 4.2 and Figure 4.3 correspond respectively to the results on Accuracy in Table A.1 and on Balanced Accuracy in Table A.2, both reported in Appendix A. Each colored line represents the performance on a specific restored version of the original data. Original data performance is plotted as the azure line.

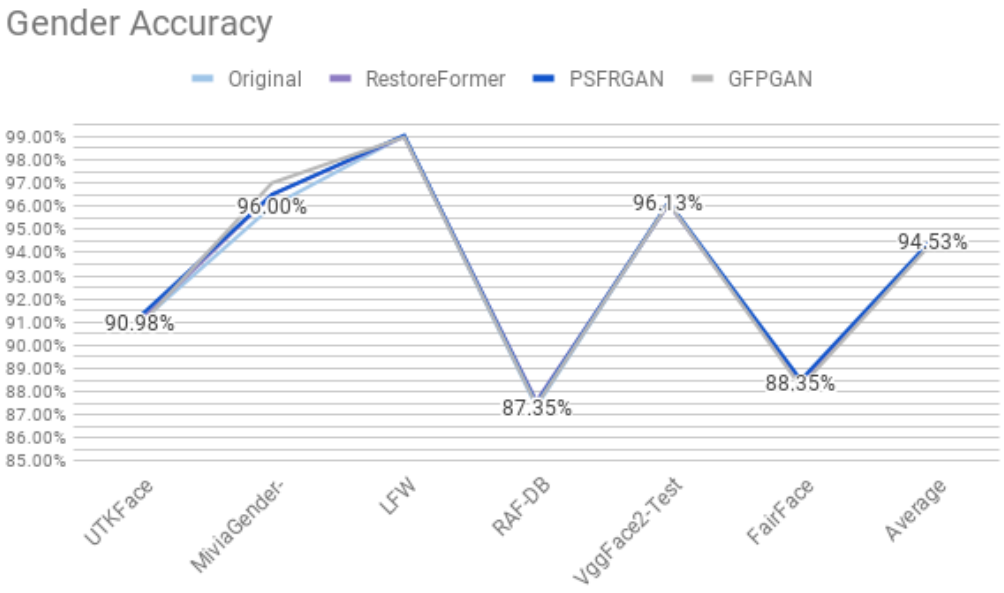

Figure 4.2: Gender Classification Accuracy.

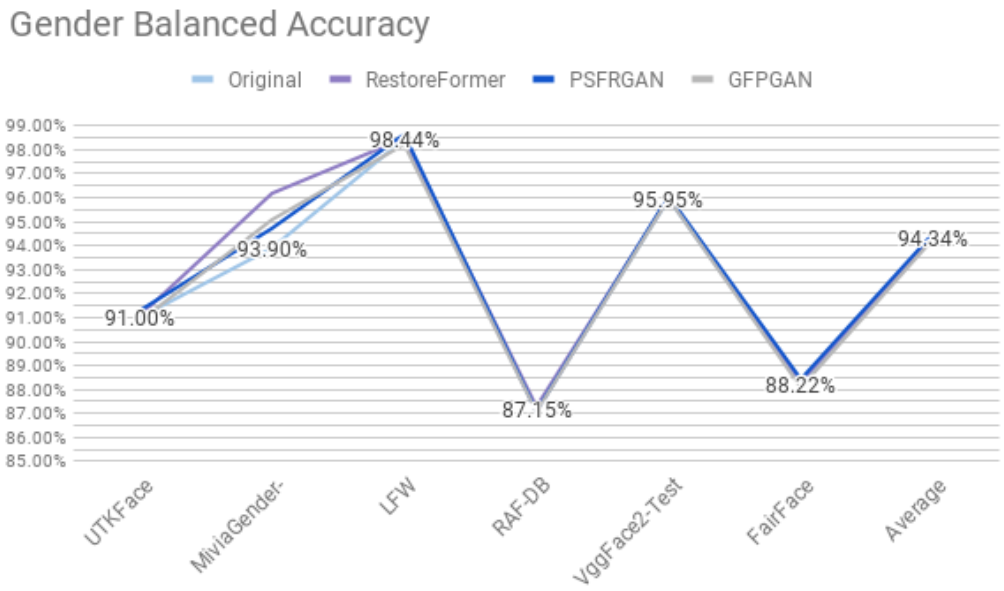

Figure 4.3: Gender Classification Balanced Accuracy.

The plot shows that all the lines are very close to each other, almost overlapping, meaning that Blind Face Restoration, in general, doesn't influence the performance much. For this reason, it is possible to deduce that, at least, **the restoration process is able to preserve gender information**.

Figure 4.4 and Figure 4.5, instead, correspond to the results of Table A.3 and Table A.4 respectively, and they're both reported in Appendix A. Those plots aim to represent the degradation or the improvement of the performance with a $\Delta score$ value, in this case, ΔAccuracy and ΔBalanced Accuracy:

$$\Delta score(\mathcal{D}, v) = score(\mathcal{D}, v) - score(\mathcal{D}, O) \tag{4.1}$$

where $score(\mathcal{D}, v)$ is the score on the dataset $\mathcal{D}$ (row of the table) in the restored version v (column of the table) and $score(\mathcal{D}, O)$ is the score on original version of the dataset. **Positive values of $\Delta score$ mean an increment in performance, while negative values have the opposite meaning.** With this in mind, the upper a point of a colored line is located, the better the achieved performance is. Therefore, if a line has the tendency to overtake another line, the performance on the dataset version related to the first line is, on average, better than the second one.

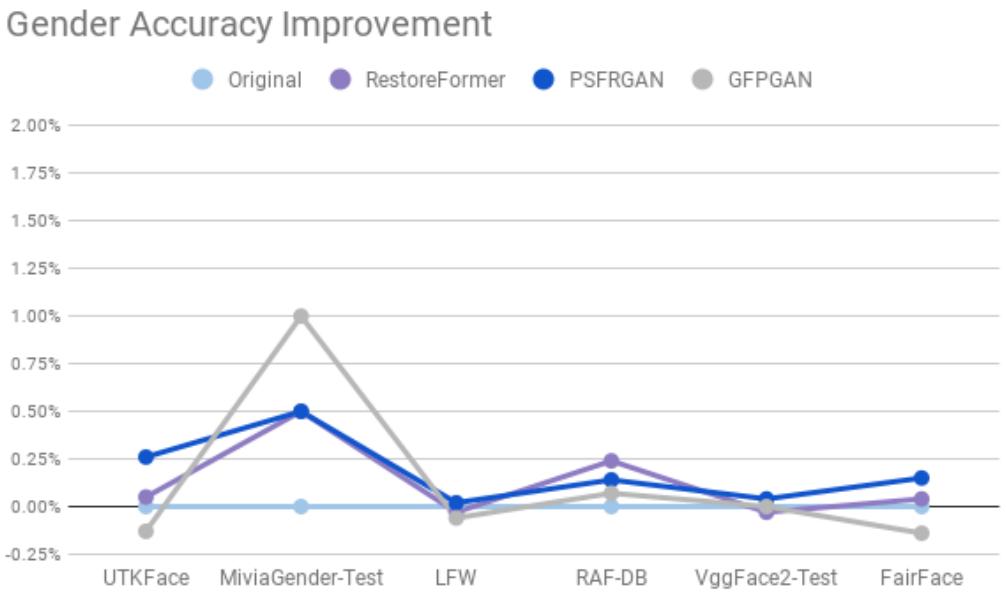

Figure 4.4: Gender Classification ΔAccuracy.

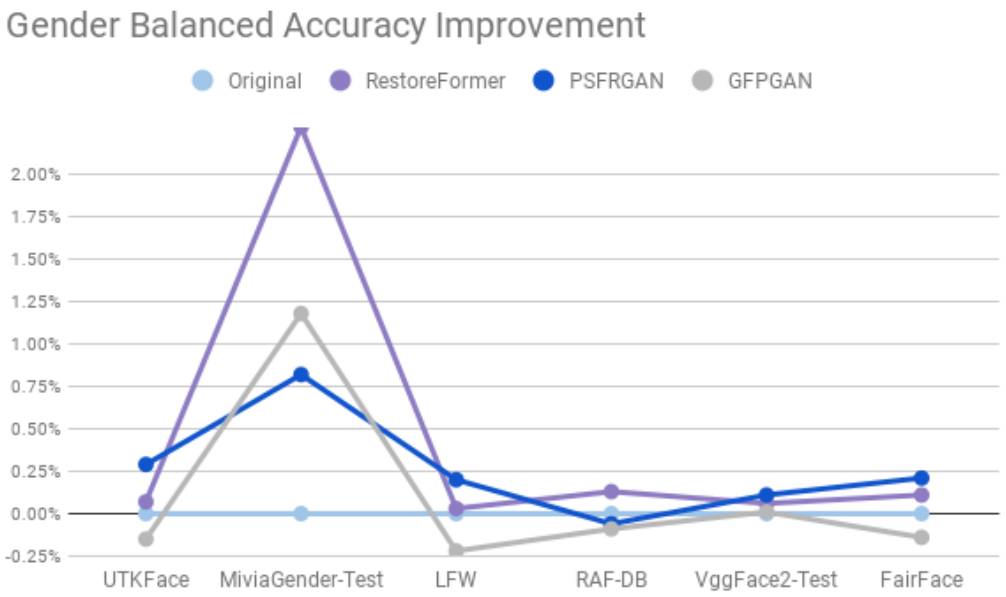

Figure 4.5: Gender Classification ΔBalanced Accuracy.

PSFRGAN provides the best results among the restoration networks, as its line tends, on average, to overtake the other networks' lines and also the Original Data line, meaning that it leads to a little increase in gender recognition performance. GFPGAN, instead, provides the worst results and a slight degradation of the scores. Both these considerations become more evident in the Balanced Accuracy score.

To check the correlation between Blind Face Restoration and the gender's traits preservation, the conducted experiment analyzes the rate of occurrence f for each class c in the predictions p. The rates calculated are averaged on all the available datasets for the task. As for the improvement/degradation plot, Δf is considered to measure the variation of prediction frequency for each class as follows:

$$f(c,p) = \frac{n_{c,p}}{n_{tot}} \tag{4.2}$$

$$\Delta f(c,p) = f(c,p) - f(c,l) \tag{4.3}$$

where $n_{c,p}$ is number of samples with class c in predictions p and n_{tot} is the total number of samples. Predictions p can be the ones inferenced from Original data

or from a restored version of them. $f(c, l)$ is the rate of samples with class c in real labels l.

Positive $\Delta f(c, p)$ values mean that the number of predictions for a certain class c is increased with respect to the Original labels. The scope is to find out whether the number of predictions for a certain class increase, maybe as a projection of the bias inherited by the unbalanced Blind Face Restoration networks' training set, FFHQ. Figure 4.6 show the variation in the prediction rate Δf for both classes *male* and *female* as a bars plot. Δf calculation follows the same principle as $\Delta score$. **Therefore, a bar growing above the abscissa axis means an increment of predictions for that class with respect to the Original data's predictions. A bar growing down the axis has the opposite meaning**. Class rates for all soft biometrics are available in Table A.19 and the corresponding variation values are available in Table A.20, both in Appendix A.

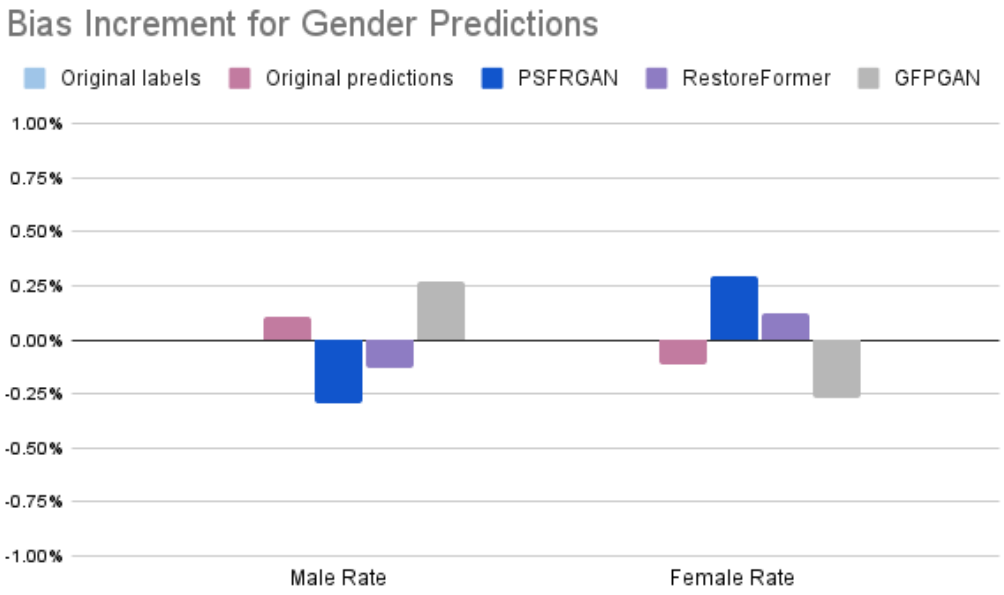

Figure 4.6: Gender $\Delta f(c, p)$ for each class c and predictions p.

According to the analysis, there is no evident impact on predictions, with such increment oscillating approximately between the very short interval $[-0.25\%, 0.25\%]$. The bias inherited by the model used for the predictions

is also very small. This is probably due to the high availability of both male and female face pictures on the web, from which FFHQ has been extracted. For this reason, the restoration networks may be able to generalize well on such soft biometric and successfully preserve gender information after the reconstruction. It also leads, sometimes, to a slight improvement in performance, even without the need for a gender-preserving loss function in training.

4.4.2 Age Estimation

The results of the age estimation task are retrieved with age SENet. Figure 4.7 reports the Mean Absolute Error values of Table A.5 in Appendix A.

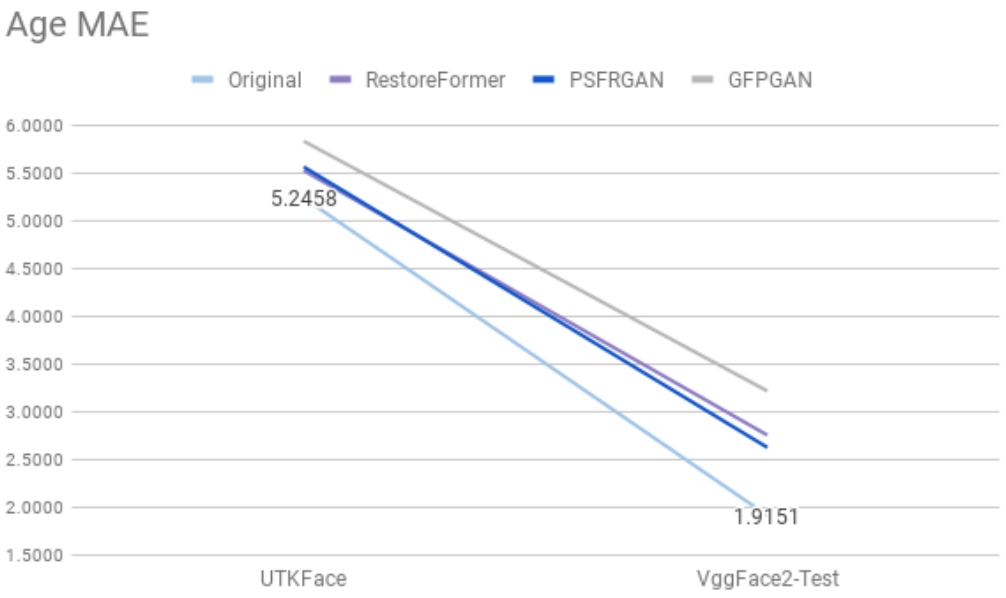

Figure 4.7: Age Mean Absolute Error

Original data provides the lowest MAE score meaning that the Blind Face Restoration degrades performance on age estimation for all the tested networks. Figure 4.8 shows the degradation of each restoration with respect to the Original data. The values are reported in Table A.6 in Appendix A. For better visualization of the results, $\Delta score$ is negated

this time, because the performance increases when the MAE decreases. This way, the plot still represents the colored lines with the best performance above the other ones.

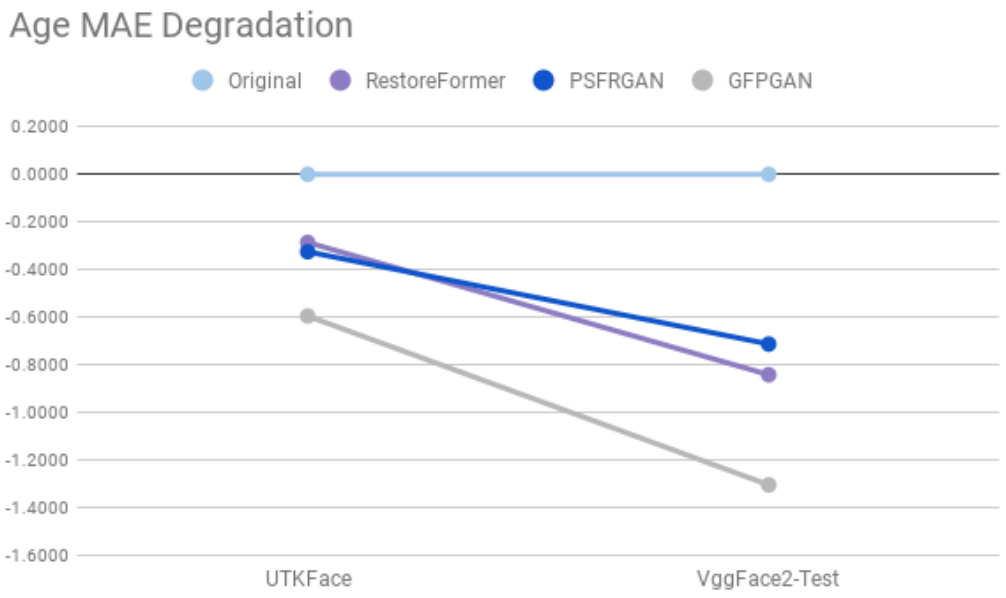

Figure 4.8: Age ΔMAE

The Original data line located on the abscissa axis confirms to be the upper one. Among the restoration networks, PSFRGAN gives, on average, the fewest degradation, with its line being the closest to the axis. GFPGAN has, again, the worst performance with its line located under the other ones. Such degradation is less impacting on UTKFace, which contains in-the-wild gathered faces. Going in-depth, RestoreFormer also provides the slightest reduction on UTKFace.

The analysis of the restoration effect on age traits focuses on the distribution found in the predictions. Figure 4.9 reports mean variation $\Delta\mu_p$ and standard deviation variation $\Delta\sigma_p$ for each prediction set p with respect to the original labels l as it was for gender classification.

$$\Delta\mu_p = \mu_p - \mu_l \tag{4.4}$$

$$\Delta\sigma_p = \sigma_p - \sigma_l \tag{4.5}$$

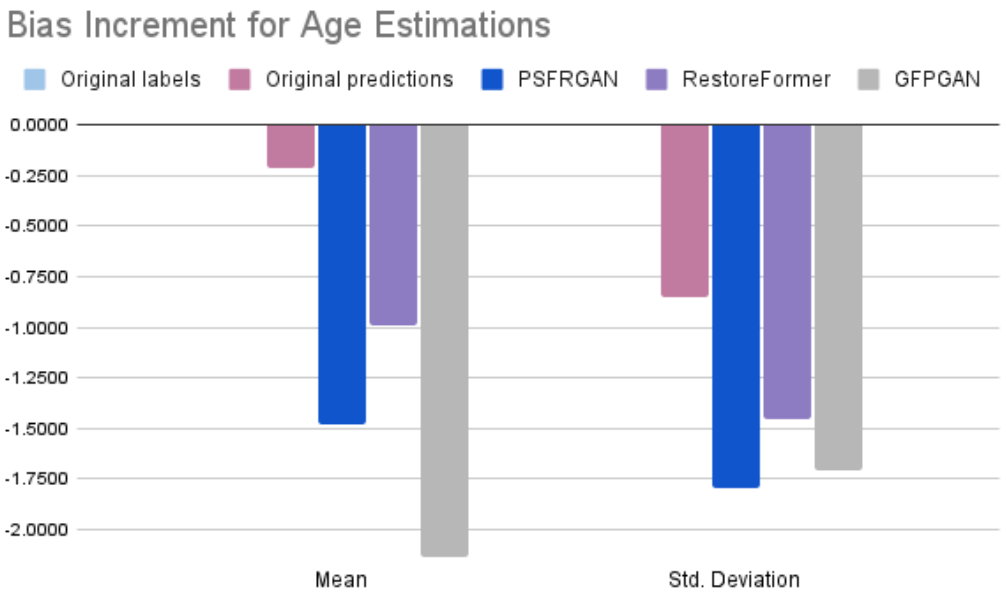

Figure 4.9: Age's mean $\Delta\mu_p$ and variance $\Delta\sigma_p$ relative to Original data predictions.

Bars in Figure 4.9 grow downward meaning that both the average predicted age and the standard deviation decrease after the restoration, with respect to the Original data.

A left shift of the mean predicted age means that the restoration networks are likely to have a rejuvenating effect on the faces. This is evident in the qualitative analysis conducted in Section 4.1. The reconstruction process has a clear softening effect on wrinkles, especially with GFPGAN, which has, in fact, the highest mean shift and also the lowest performance on this task.

A left shift of the standard deviation in predicted age distributions means that the face reconstruction tends to shift the age traits toward the mean value. This may depend on the vast majority of middle-aged samples, being FFHQ gathered in controlled conditions. Limit cases like very young children or very old people appear to be rarer. In other words, Blind Face Restoration tends, even more, to rejuvenate old age faces and tends to age very young ones. Part of this behavior is addressable to the

age estimation model, already introducing an almost negligible, yet present, negative contribution. PSFRGAN has the highest decrease in standard deviation. However, despite the difference being considerable with respect to the Original Labels and the Original predictions, it becomes more and more negligible when compared to the other Blind Face Restoration networks.

No considered Blind Face Restoration network takes into account an age-preserving loss function. Differently from gender, age information is more biased, and borderline cases are less available. Due to this choice, such networks may have developed their typical softening property which is moving the average predicted age on the left.

4.4.3 Ethnicity Classification

The results of the ethnicity classification task are retrieved with VggFace. Figure 4.10 reports the Accuracy values of Table A.7 and Figure 4.11 reports the Balanced Accuracy values of Table A.8, both available in Appendix A.

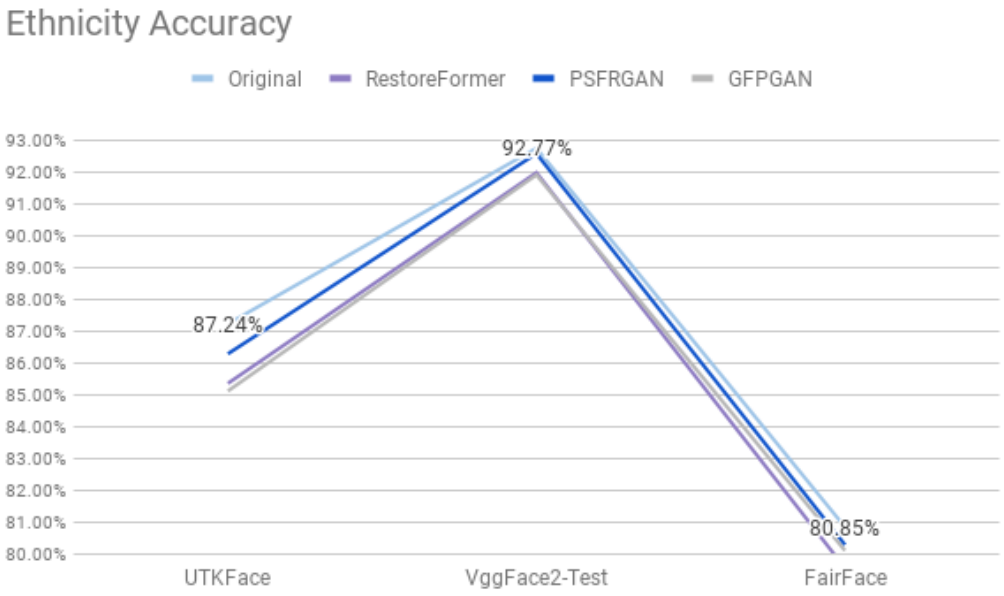

Figure 4.10: Ethnicity Classification Accuracy

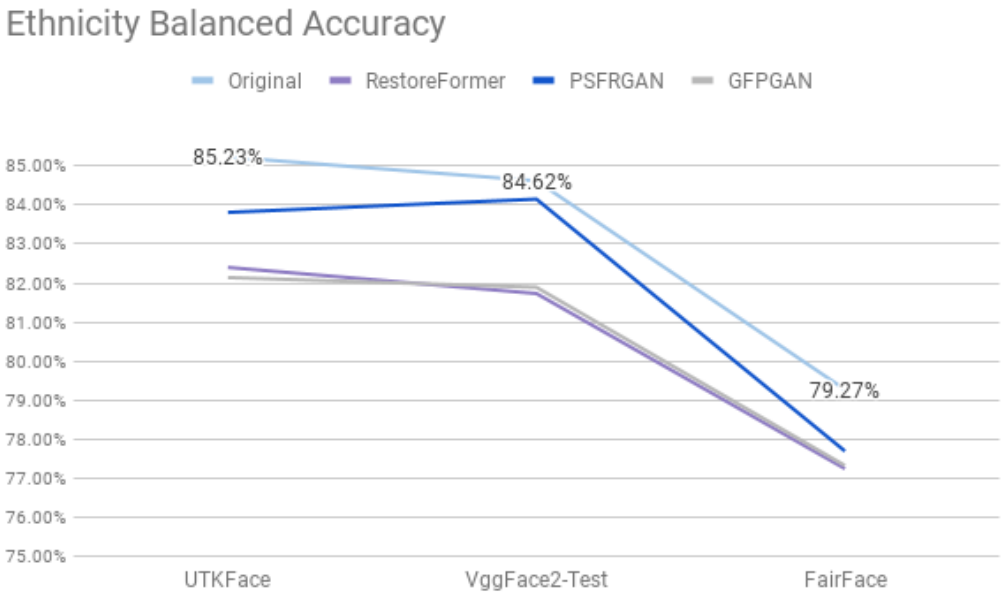

Figure 4.11: Ethnicity Classification Balanced Accuracy

The Original Data line is visibly above the other, therefore the Blind Face Restoration causes a non-negligible drop in performance for ethnicity classification. Following the same principle of plot in Section 4.4.1, Figure 4.12 shows the distribution of ΔAccuracy reported in Table A.9, Appendix A, among the datasets provided with ethnicity labels. Figure 4.13 does the same for ΔBalanced Accuracy values in Table A.10, Appendix A.

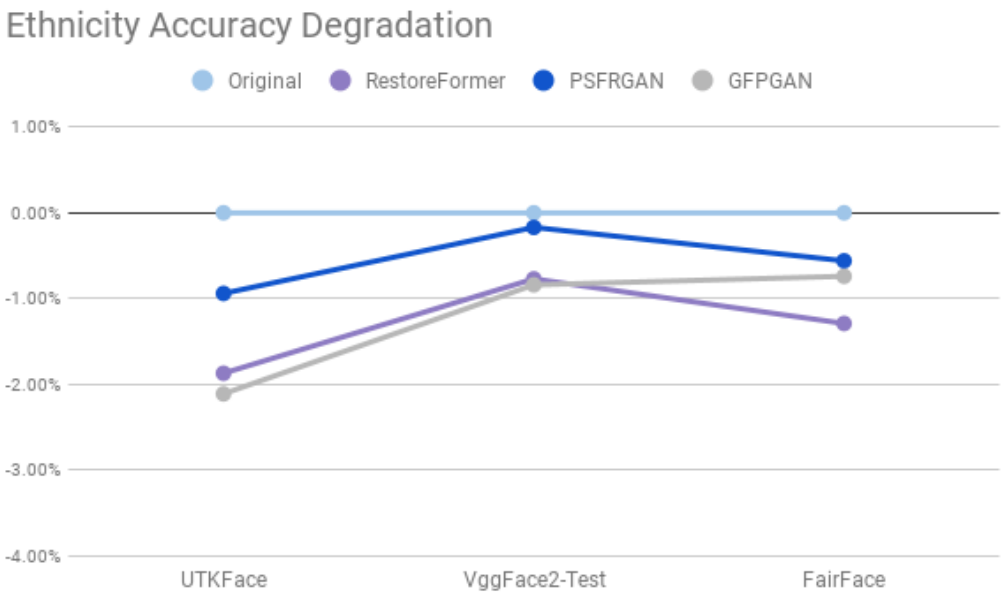

Figure 4.12: Ethnicity Classification ΔAccuracy

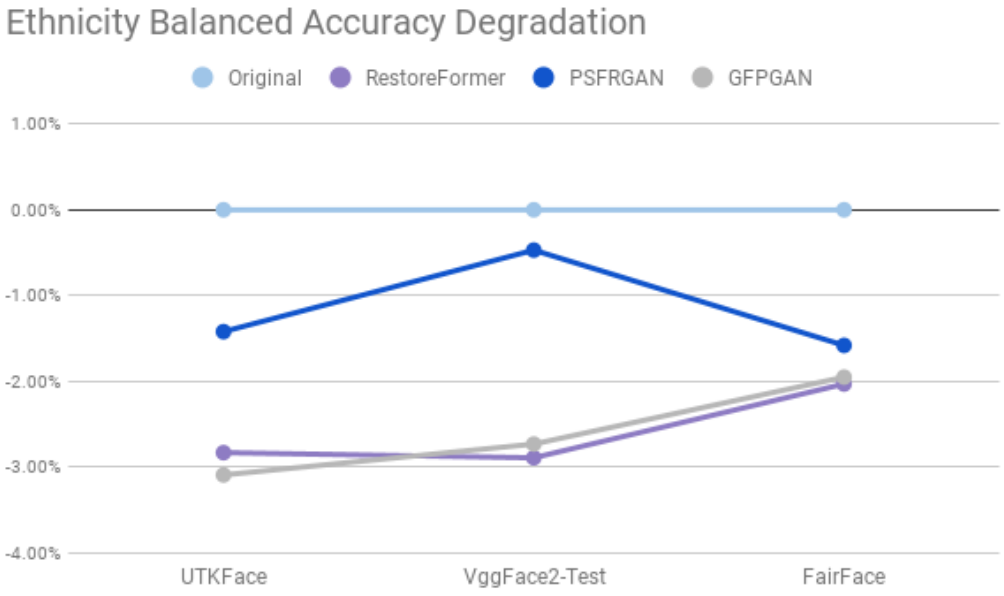

Figure 4.13: Ethnicity Classification ΔBalanced Accuracy

This degradation of performance is far more evident in the Balanced Accuracy score where RestoreFormer and GFPGAN both cause a decrease of around the 3%. Considering Accuracy, which is more affected by bias effects,

performance on UTKFace and VggFace2 Test are better than FairFace by a large amount. The way FairFace differs from them lies in its data distribution, which is more balanced among the classes. While all the datasets, share in common a vast majority of Caucasian Latin samples, VggFace2 Test, in particular, has a very large Caucasian Latin rate of 76.24%. Figure 4.14 reports the FFHQ dataset's distribution for ethnicity labels according to the study of Maluleke et al. [32].

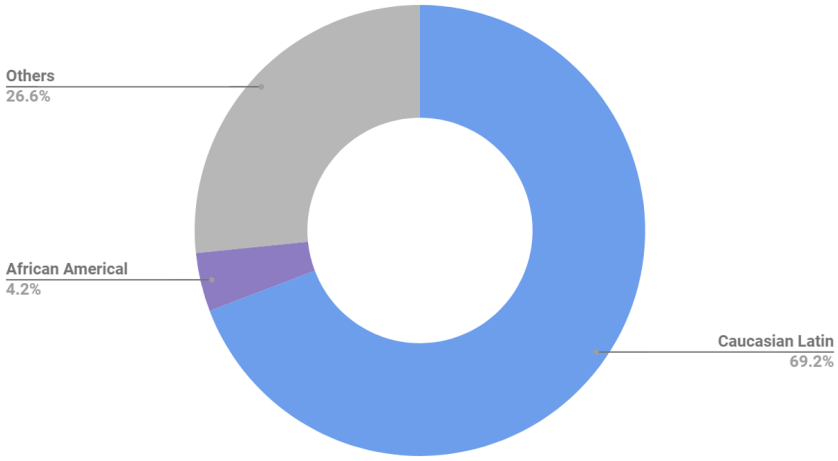

Figure 4.14: FFHQ ethnicity distribution.

Caucasian Latin samples are predominant by the far amount of 69.2%. Continuing to follow the protocol already adopted in Section 4.4.1, Figure 4.15 shows the class frequency variation in the predictions after the restoration.

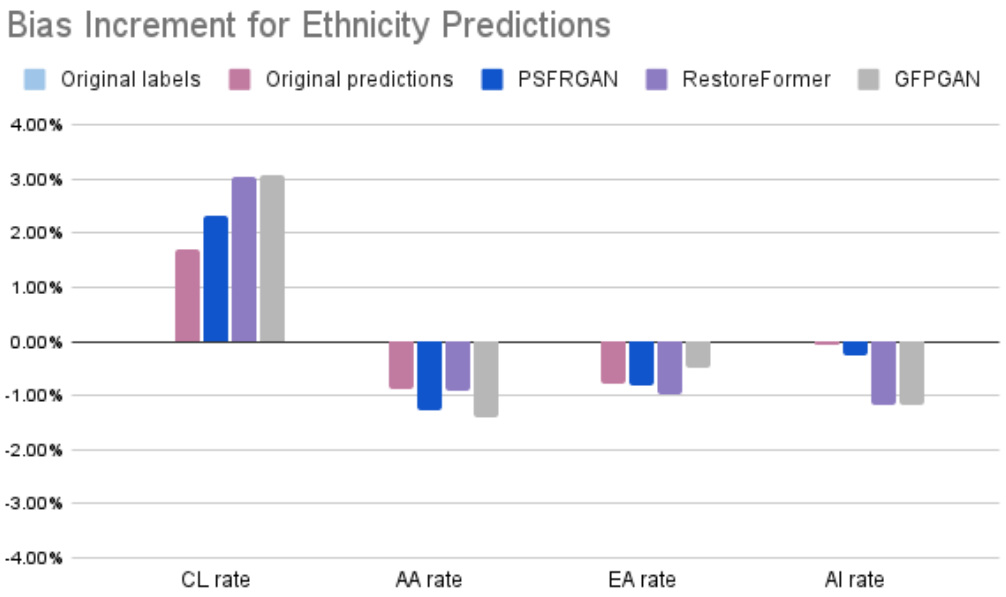

Figure 4.15: Ethnicity $\Delta f(c, p)$ for each class c and predictions p.

As inferable from Figure 4.14 distribution and confirmed by the rate bars, the Caucasian Latin predictions are the only ones growing while the other classes are affected by a visible reduction. Blind Face Restoration networks inherit FFHQ bias and generate reconstructed faces with traits getting more similar to Caucasian Latin after the restoration. This is a possible explanation for why the performance on Accuracy score is largely superior on VggFace2 Test which has the largest rate of Caucasian Latin among the datasets.

As for gender and age, there's no ethnicity-preserving loss function involved in the training of the Blind Face Restoration networks. Differently from gender classification, ethnicity classification is a multi-class problem with a high unbalance of classes in available data. Training based on such conditions is ill-posed and inherits data bias as previously proved.

4.4.4 Facial Emotion Classification

The results on the ethnicity classification task are retrieved with Multitask Seresnet B. Figure 4.16 and Figure 4.17 plot, respectively, Accuracy values in Table A.11 and Balanced Accuracy values in Table A.12, both

available in Appendix A. The chosen representation is a bar diagram, given the availability of only one dataset provided with facial emotion labels. In this representation, the higher a colored bar is, the better the performance on the related version of the dataset is. Original Data are still represented with the azure color.

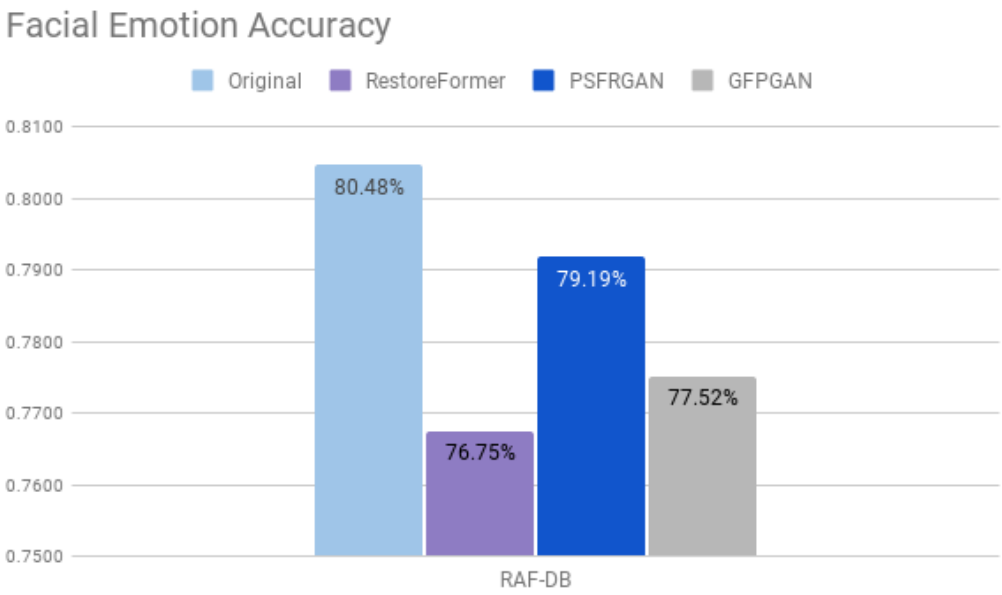

Figure 4.16: Facial Emotion Classification Accuracy

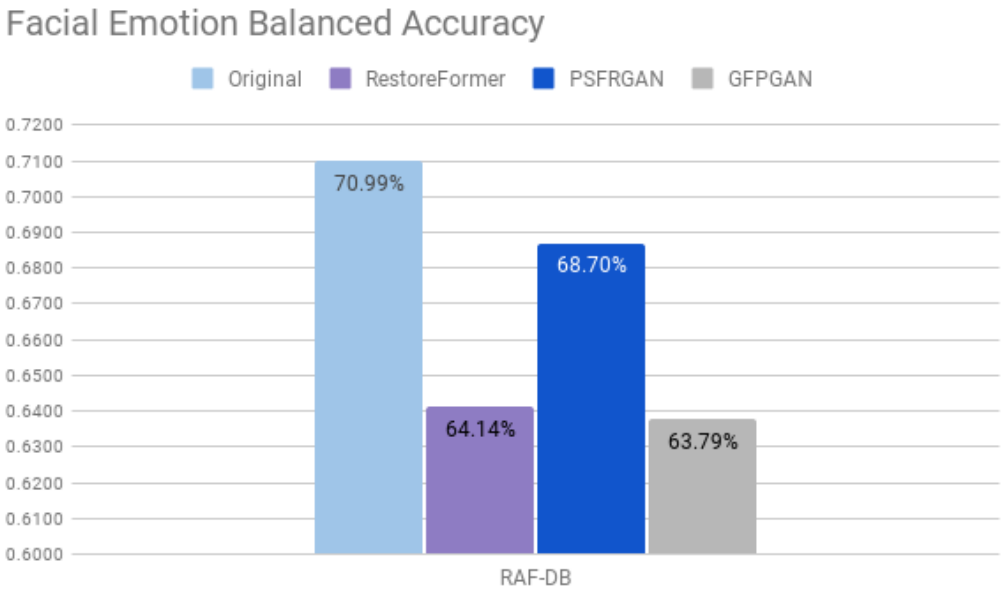

Figure 4.17: Facial Emotion Classification Balanced Accuracy

The Original data bar is far higher than the others, meaning that Blind Face Restoration causes a heavy decrease in performance on facial emotion recognition task. Among all the considered soft biometrics, this task is the hardest and provides not only the worst performance but also the most severe drop in Accuracy and Balanced Accuracy scores. Figure 4.18 and Figure 4.19 plot, respectively, Accuracy values in Table A.13 and Balanced Accuracy values in Table A.14, both available in Appendix A.

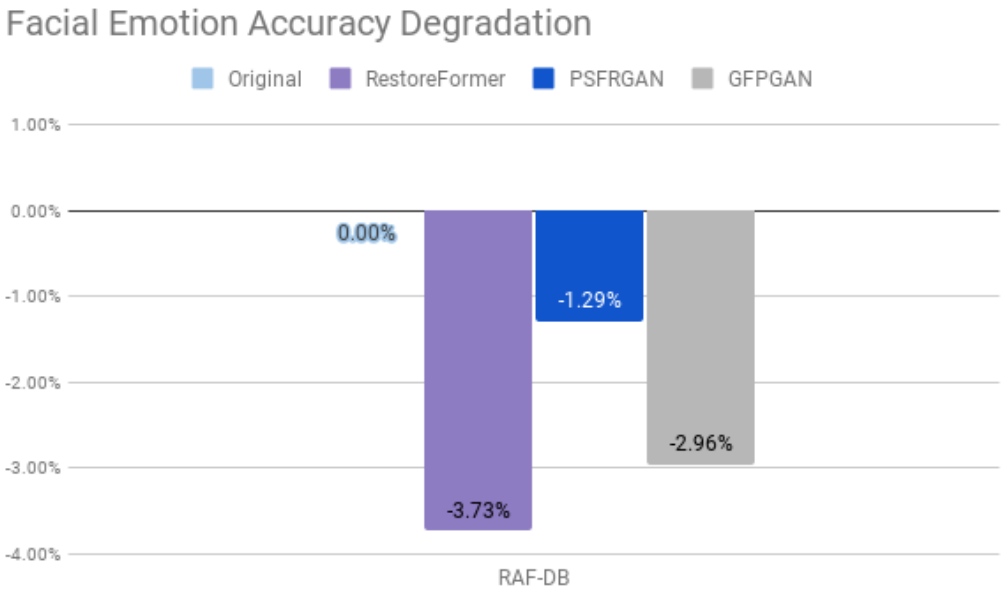

Figure 4.18: Facial Emotion Classification ΔAccuracy

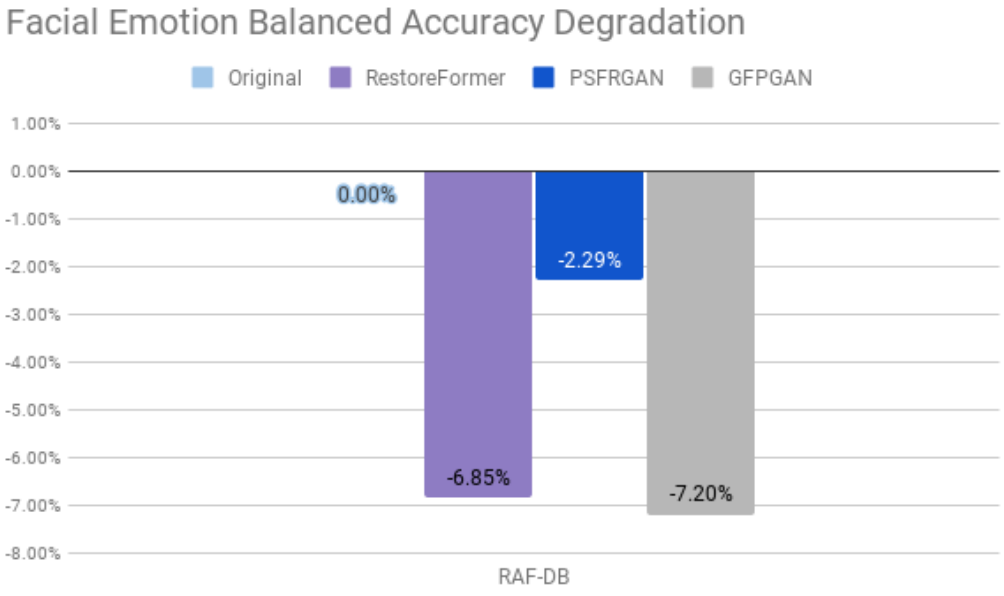

Figure 4.19: Facial Emotion Classification ΔBalanced Accuracy

Colored bars are all growing downward confirming that the introduced variation of performance is a negative contribution. This decrease, as for the previous soft biometrics, becomes more evident on the Balanced Accuracy

score, where GFPGAN reaches a heavy drop of 7.20%. RestoreFormer provides a negative contribution only slightly better than GFPGAN, while, instead, PSFRGAN is still the best soft biometrics preserving network, with a very small reduction of -2.29% compared to the others.

The reason behind this behavior, as for ethnicity classification and gender estimation, can be found behind the ill-posed training of the models, as confirmed by Figure 4.20.

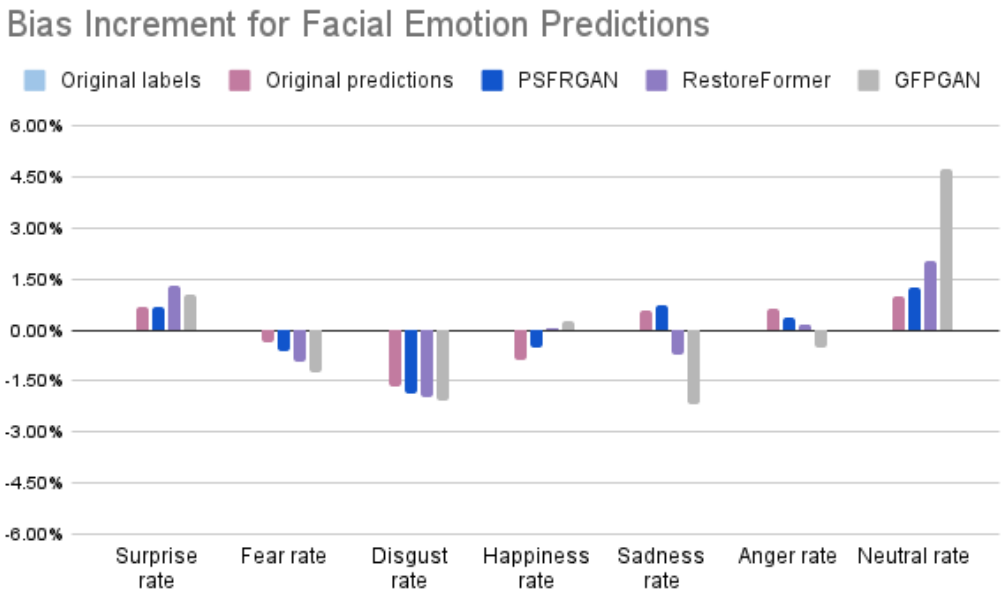

Figure 4.20: Ethnicity $\Delta f(c, p)$ for each class c and predictions p.

According to the analysis, all the models, with the inclusion of a little contribution by the classifier itself, have a strong tendency to neutralize facial emotions. This is visible on the Neutral rate bars growing up much higher than the others. Also, the number of predicted Surprised faces gets a little increase. Fear, Disgust, and Sadness rate of appearance in the predictions is reduced, because such emotions are harder to find in pictures taken in controlled conditions like FFHQ's. The other emotions get negligible oscillations around the Original labels rates.

As for all the previous soft biometrics, no facial-emotion-preserving loss function has been optimized during the training of the analyzed models.

Like ethnicity, this is a multi-class problem with, possibly, an even higher unbalance of classes in available data. For the same reason behind ethnicity bias inheritance, restoration networks are also affected by the facial emotion bias of the training set.

4.4.5 Identification

The results of the identification task are retrieved with the VggFace descriptor and the procedure presented in Section 3.4. Figure 4.21 reports the Accuracy values of Table A.15 and Figure 4.22 reports the Balanced Accuracy values of Table A.16, both available in Appendix A. As for Facial Emotion, the bar plots are chosen as they are more suitable when there's only one dataset provided with identity labels.

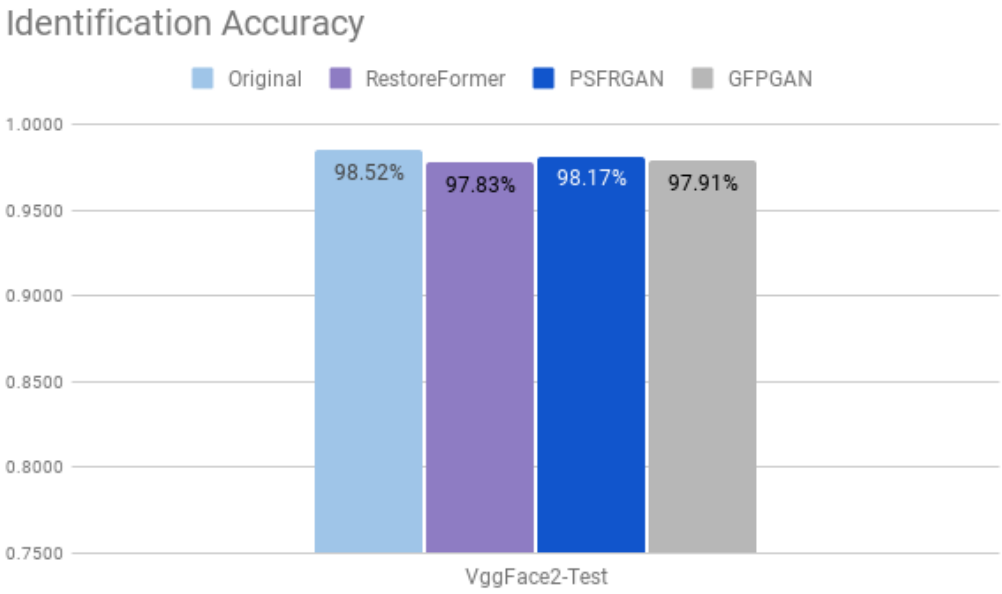

Figure 4.21: Identification Classification Accuracy

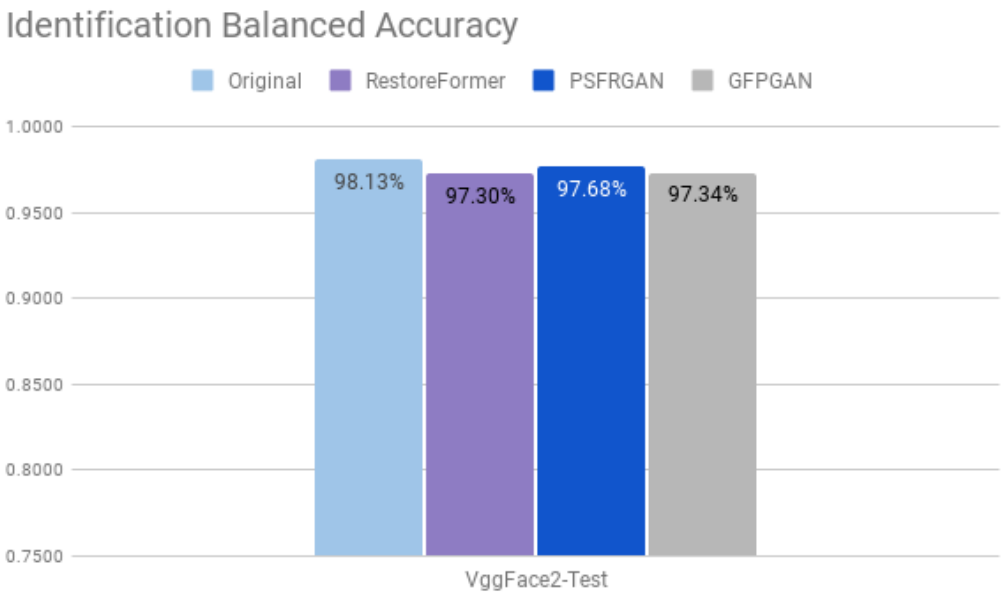

Figure 4.22: Identification Classification Balanced Accuracy

Differently from Ethnicity and Facial Emotion cases, the performance of identification is almost the same on all the dataset's versions. However, Accuracy and Balanced Accuracy scores are still better before the restoration, even if only by a very small margin. Figure 4.23 and Figure 4.24 plot, respectively, ΔAccuracy values in Table A.17 and ΔBalanced Accuracy values in Table A.18, both available in Appendix A.

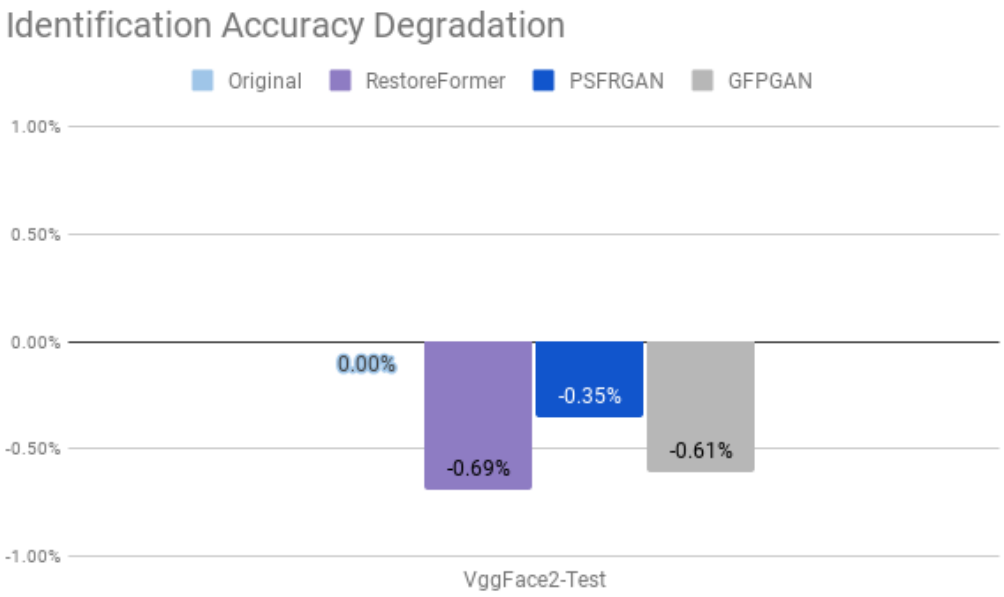

Figure 4.23: Identification Classification ΔAccuracy

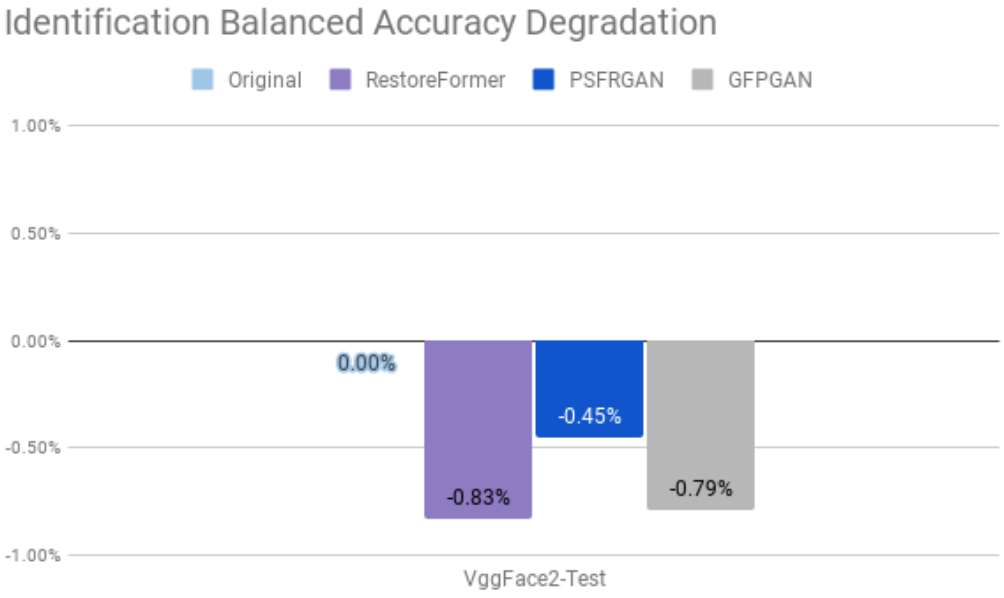

Figure 4.24: Identification Classification ΔBalanced Accuracy

The decrease does not exceed the value of 0.83%, therefore, it's almost negligible. However, even though the performance is very close on all the restored versions of the dataset, PSFRGAN keeps being the best preserving

model while RestoreFormer and GFPGAN still settle themselves around the same values. Differently from the other soft biometrics, FFHQ is not suitable for a bias analysis of identity information. Anyway, a different analysis can be conducted, focusing the attention on the loss functions involved during training.

Both GFPGAN and RestoreFormer provide an identity-preserving loss function optimized on the ArcFace network. Furthermore, all the considered networks, PSFRGAN included, take into account a perceptual loss that optimizes the preservation of the features extracted with VGG-19. This may explain why performance isn't dropping after the restoration as it happens with age, ethnicity, and facial emotion. Although this contribution has a visible preserving effect on the performance, the networks still fail to properly generalize on identity, leading to a small, yet noticeable, decrease.

CHAPTER 5

CONCLUSIONS

This thesis describes the design and implementation of an experimental framework to test the eventual impact of Blind Face Restoration on soft biometrics recognition tasks. It also carries the experiment on with three restoration networks (RestoreFormer, PSFRGAN, GFPGAN) and seven datasets (CelebA-HQ, VggFace2 Test, UTKFace, LFW, MiviaGender Test, RAF-DB Test, and FairFace Test) gathered in different conditions. The objective is to measure the restoration performance in terms of elaboration speed and image reconstruction quality. However, the test mainly aims to gather, compare, and analyze the scores on soft biometrics recognition tasks after the restoration, to find correlations behind the reconstruction process and such performance.

According to the results, all the Blind Face Restoration networks seem to be very slow, even on a competitive GPU asset like the one involved in the experiment. This means that real-time specification, mandatory in tasks like video surveillance, may be hard to satisfy without high-end expensive hardware.

Regarding the quality of the reconstructed faces, Image Quality Metrics report good values on all the networks, especially on GFPGAN which is able

to provide the best restoration with the best elaboration time. Such primacy in terms of raw and human-perceived quality is confirmed by the qualitative analysis.

The most relevant conclusion about the whole experiment is that **Blind Face Restoration doesn't preserve all soft biometrics**, even with a good restoration quality. The performance achieved by soft biometrics recognition networks on the same test data, after the restoration process, tends to have a non-negligible drop. As explored with an analysis focused on the training of the networks, such models are affected by two most important conditions: their predictions are ill-posed and project the data unbalance found in the training set FFHQ, and they do not take into account soft-biometrics-preserving losses. With this in mind, reconstructed faces are more likely to look like the faces seen by the networks during the training face. Ethnicity distinctive features become more similar to Caucasian Latin features because they are the majority in the training set and, for the same reason, facial expressions get neutralized and age gets nearer to the mean.

Another interesting observation regards GFPGAN, which is the best network for the quality of the restoration but has, on average, the worst degrading effect on soft biometrics recognition performance.

5.1 Future Developments

The experimental framework is designed to be extended and many ideas can be included in a possible continuation of the experiment:

- Blind Face Restoration's most used training set in the state of the art is FFHQ. It provides many high-quality faces but it's very unbalanced on soft biometrics, contaminating the reconstruction with its inherited bias. A good test to check if it is possible to overcome the bad effect of restoration on soft biometrics recognition is to re-train the Blind Face Restoration models. Providing a new less-biased high-quality training set could generate models which can eventually preserve soft biometrics

better, not being ill-posed as the current state-of-the-art solutions;

- The current Blind Face Restoration networks don't consider soft-biometrics-preserving loss functions, except for the identity in which, in fact, the performance drop is very small. As a possible future addition, some soft biometrics recognition single-task or multi-task networks can be introduced into the training of the restoration models, along with an objective function optimized for soft biometrics preservation. This may have a positive effect, reducing the negative impact of the restoration on soft biometrics recognition, or even improving its performance;

- The framework may be extended even more to have more reliable results by, for example, adding new datasets, new Blind Face Restoration frameworks, and new soft biometrics recognition networks into the experiment.

ACKNOWLEDGEMENT

This thesis is the end of a long academic journey, which origin I like to place on the very first day I picked up a pen in primary school. I've always walked this path with passion and found friendship, satisfaction, and hard work on my way. This master's degree doesn't mark an end to my studies because, as a popular character from one of my favorite video-game says, "a true master is an eternal student". Nonetheless, this still is a significant goal in my progression, and I can't forget anyone of the important people that, aware of this or not, have contributed to making me the person I'm today.

I want to thank professor Antonio Greco, my supervisor, which followed me in this thesis work with extreme patience and great skills. He has also been a recurring reference point during my academic journey and taught me so much during his academic lectures.

I also wanted to thank professor Mario Vento, my co-supervisor who has followed my thesis work behind the scenes, always providing feedbacks and ensuring that everything went according to plan. Working under his supervision has been a source of great pride for me.

I want to thank each professor that, during my entire academic journey, was able to pass me their knowledge but, most importantly, their passion for the matter.

I am thankful to all my colleagues, who have, with their kindness

and intellectual vivacity, contributed to creating a positive, competitive, and stimulating learning environment. I want to thank, in particular, Gerardo Bottiglieri and Simona Sorgente who went through this amazing journey always by my side, becoming much more life companions than academic friends. I must also dedicate part of my gratitude to anyone who, in the hard moments of this journey, helped me, without being aware, to overcome them and change for the best.

I infinitely thank Stefano, my beloved cousin. Being two years older, he has always been my irreplaceable bright reference point, even a mentor sometimes. I keep the memory of each precious moment spent together during our growth jealously in my heart. In the moments of blind uncertainty, when everything seemed dark, I always knew that thanks to him, I could see at least a few steps ahead. I hope we'll continue to support each other during our lives, as we've always done.

I must thank all my extra-academic friends who have always supported me in the good and bad times, always giving me the charge to face difficulties and hard work. Every single one of them gave a significant contribution to who I'm today. They are indisputable proof that family can be chosen. I must dedicate a special thanks to Antonio Brienza, my eternal friend who, after all those years, continues to remind me how our friendship is unwavering, no matter the circumstances.

I want to thank aunt Concetta, uncle Mario, Marco, aunt Liana, and every one of my in-laws and blood relatives, who have always stood by my side, caring for me and showing me love at any stage of my life, unconditionally.

I'm extremely thankful to my grandmothers Tina and Pasqualina, who passed away in recent years. The one thing they share in common is the immeasurable pride and love they had for me, and I hope to always be worthy of that. With them, I want to thank also everyone who cared for me and is not here anymore to rejoice with me in this success.

Finally, my most heartfelt thanks go to my mother Catia, and my father Luigi. No matter how much I try, there are no words capable of properly

expressing my infinite gratitude for what they have done for me, and there's nothing I can possibly do to pay back the effort they put into my raising. With this in mind, I know the only thing I can do is continue to be worthy of their unquestioned love, day after day. In particular, I want to thank them for having thought me to be curious and passionate in everything I do, and for the moral principles they passed me, that will always be my compass in my moments of doubt.

In conclusion, I hope to continue to deserve all the love and support I constantly receive in my everyday life. This achievement ends the chapter of a book that I hope will continue to fill itself with special people and exciting experiences for a long time, and I really wish the same luck to everyone else.

REFERENCES

[1] Xintao Wang et al. "Towards Real-World Blind Face Restoration With Generative Facial Prior". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 9168–9178.

[2] Chaofeng Chen et al. "Progressive Semantic-Aware Style Transformation for Blind Face Restoration". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 11896–11905.

[3] Xiaoming Li et al. "Enhanced Blind Face Restoration With Multi-Exemplar Images and Adaptive Spatial Feature Fusion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[4] Xiaoming Li et al. "Learning Warped Guidance for Blind Face Restoration". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. Sept. 2018.

[5] Xiaoming Li et al. "Blind Face Restoration via Deep Multi-scale Component Dictionaries". In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi et al. Cham: Springer International Publishing, 2020, pp. 399–415. ISBN: 978-3-030-58545-7.

[6] Zhouxia Wang et al. “RestoreFormer: High-Quality Blind Face Restoration From Undegraded Key-Value Pairs”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 17512–17521.

[7] Tao Yang et al. “GAN Prior Embedded Network for Blind Face Restoration in the Wild”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2021, pp. 672–681.

[8] Jia Li et al. “Universal Face Restoration With Memorized Modulation”. In: *arXiv preprint arXiv:2110.01033* (2021).

[9] Yu Liu. “Face Restoration Network with Feature Prior”. In: *2021 IEEE International Conference on Computer Science, Artificial Intelligence and Electronic Engineering (CSAIEE)*. IEEE. 2021, pp. 222–226.

[10] Domonkos Varga. “No-reference image quality assessment with global statistical features”. In: *Journal of Imaging* 7.2 (2021), p. 29.

[11] Marius Pedersen, Jon Yngve Hardeberg, et al. “Full-reference image quality metrics: Classification and evaluation”. In: *Foundations and Trends® in Computer Graphics and Vision* 7.1 (2012), pp. 1–80.

[12] Richard Zhang et al. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.

[13] Martin Heusel et al. “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. In: *Advances in neural information processing systems* 30 (2017).

[14] Min Jin Chong and David Forsyth. “Effectively Unbiased FID and Inception Score and Where to Find Them”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020.

[15] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. "No-reference image quality assessment in the spatial domain". In: *IEEE Transactions on image processing* 21.12 (2012), pp. 4695–4708.

[16] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. "Making a "completely blind" image quality analyzer". In: *IEEE Signal processing letters* 20.3 (2012), pp. 209–212.

[17] N Venkatanath et al. "Blind image quality evaluation using perception based features". In: *2015 Twenty First National Conference on Communications (NCC)*. IEEE. 2015, pp. 1–6.

[18] Antitza Dantcheva et al. "Bag of soft biometrics for person identification". In: *Multimedia Tools and Applications* 51.2 (2011), pp. 739–777.

[19] Antonio Greco et al. "Gender recognition in the wild: a robustness evaluation over corrupted images". In: *Journal of Ambient Intelligence and Humanized Computing* 12.12 (2021), pp. 10461–10472.

[20] Antonio Greco et al. "Benchmarking deep networks for facial emotion recognition in the wild". In: *Multimedia Tools and Applications* (2022), pp. 1–32.

[21] Antonio Greco et al. "Effective training of convolutional neural networks for age estimation based on knowledge distillation". In: *Neural Computing and Applications* (2021), pp. 1–16.

[22] Antonio Greco et al. "Benchmarking deep network architectures for ethnicity recognition using a new large face dataset". In: *Machine Vision and Applications* 31.7 (2020), pp. 1–13.

[23] Pasquale Foggia et al. "Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition". In: *Engineering Applications of Artificial Intelligence* (2022).

[24] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. "Deep Face Recognition". In: *British Machine Vision Conference*. 2015.

[25] Tero Karras et al. “Progressive growing of gans for improved quality, stability, and variation”. In: *arXiv preprint arXiv:1710.10196* (2017).

[26] Qiong Cao et al. “Vggface2: A dataset for recognising faces across pose and age”. In: *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE. 2018, pp. 67–74.

[27] Zhifei Zhang, Yang Song, and Hairong Qi. “Age progression/regression by conditional adversarial autoencoder”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5810–5818.

[28] Gary B Huang et al. “Labeled faces in the wild: A database forstudying face recognition in unconstrained environments”. In: *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*. 2008.

[29] Shan Li, Weihong Deng, and JunPing Du. “Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2852–2861.

[30] Kimmo Karkkainen and Jungseock 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.

[31] Wei Liu et al. “Ssd: Single shot multibox detector”. In: *European conference on computer vision*. Springer. 2016, pp. 21–37.

[32] Vongani Hlavutelo Maluleke et al. “Studying Bias in GANs through the Lens of Race”. In: 2022.

LIST OF FIGURES

LIST OF TABLES

Appendices

APPENDIX A

Table A.1: Gender classification accuracy score.

	Accuracy ↑			
Dataset	Original	RestoreFormer	PSFRGAN	GFPGAN
UTKFace	90.98%	91.03%	91.24%	90.85%
MiviaGender-Test	96.00%	96.50%	96.50%	97.00%
LFW	99.03%	99.00%	99.05%	98.97%
RAF-DB	87.35%	87.59%	87.49%	87.42%
VggFace2-Test	96.13%	96.10%	96.17%	96.13%
FairFace	88.35%	88.39%	88.50%	88.21%
Average	94.53%	94.52%	94.62%	94.48%

Table A.2: Gender classification Balanced Accuracy score.

	Balanced Accuracy ↑			
Dataset	Original	RestoreFormer	PSFRGAN	GFPGAN
UTKFace	91.00%	91.07%	91.29%	90.85%
MiviaGender-Test	93.90%	96.18%	94.72%	95.08%
LFW	98.44%	98.47%	98.64%	98.22%
RAF-DB	87.15%	87.28%	87.09%	87.06%
VggFace2-Test	95.95%	96.01%	96.06%	95.96%
FairFace	88.22%	88.33%	88.43%	88.08%
Average	94.34%	94.41%	94.50%	94.28%

Table A.3: Gender classification variation ΔAccuracy with respect to Original Data.

	ΔAccuracy ↑		
Dataset	RestoreFormer	PSFRGAN	GFPGAN
UTKFace	0.05%	0.26%	-0.13%
MiviaGender-Test	0.50%	0.50%	1.00%
LFW	-0.03%	0.02%	-0.06%
RAF-DB	0.24%	0.14%	0.07%
VggFace2-Test	-0.03%	0.04%	0.00%
FairFace	0.04%	0.15%	-0.14%
Average	0.00%	0.09%	-0.04%

Table A.4: Gender classification variation ΔBalanced Accuracy with respect to Original Data.

	ΔBalanced Accuracy ↑		
Dataset	RestoreFormer	PSFRGAN	GFPGAN
UTKFace	0.07%	0.29%	-0.15%
MiviaGender-Test	2.28%	0.82%	1.18%
LFW	0.03%	0.20%	-0.22%
RAF-DB	0.13%	-0.06%	-0.09%
VggFace2-Test	0.06%	0.11%	0.01%
FairFace	0.11%	0.21%	-0.14%
Average	0.07%	0.16%	-0.06%

Table A.5: Age estimation Mean Absolute Error score.

	MAE ↓			
Dataset	Original	RestoreFormer	PSFRGAN	GFPGAN
UTKFace	5.2458	5.5319	5.5710	5.8411
VggFace2-Test	1.9151	2.7571	2.6281	3.2200
Average	2.7600	3.4610	3.3746	3.8849

Table A.6: Age estimation Mean Absolute Error variation ΔMAE with respect to Original Data. (The variation is negated)

	ΔMAE ↑		
Dataset	RestoreFormer	PSFRGAN	GFPGAN
UTKFace	-0.2861	-0.3252	-0.5953
VggFace2-Test	-0.8420	-0.7130	-1.3049
Average	-0.7010	-0.6146	-1.1249

Table A.7: Ethnicity classification Accuracy score.

	Accuracy ↑			
Dataset	Original	RestoreFormer	PSFRGAN	GFPGAN
UTKFace	87.24%	85.37%	86.30%	85.13%
VggFace2-Test	92.77%	92.00%	92.60%	91.93%
FairFace	80.85%	79.56%	80.29%	80.11%
Average	90.28%	89.21%	89.89%	89.16%

Table A.8: Ethnicity classification Balanced Accuracy score.

	Balanced Accuracy ↑			
Dataset	Original	RestoreFormer	PSFRGAN	GFPGAN
UTKFace	85.23%	82.40%	83.81%	82.14%
VggFace2-Test	84.62%	81.73%	84.15%	81.89%
FairFace	79.27%	77.24%	77.69%	77.32%
Average	84.21%	81.42%	83.40%	81.47%

Table A.9: Ethnicity Classification ΔAccuracy.

	ΔAccuracy ↑		
Dataset	RestoreFormer	PSFRGAN	GFPGAN
UTKFace	-1.87%	-0.94%	-2.11%
VggFace2-Test	-0.77%	-0.17%	-0.84%
FairFace	-1.29%	-0.56%	-0.74%
Average	-1.07%	-0.39%	-1.12%

Table A.10: Ethnicity Classification ΔBalanced Accuracy.

	ΔBalanced Accuracy ↑		
Dataset	RestoreFormer	PSFRGAN	GFPGAN
UTKFace	-2.83%	-1.42%	-3.09%
VggFace2-Test	-2.89%	-0.47%	-2.73%
FairFace	-2.03%	-1.58%	-1.95%
Average	-2.79%	-0.80%	-2.73%

Table A.11: Facial Emotion classification Accuracy score.

	Accuracy ↑			
Dataset	Original	RestoreFormer	PSFRGAN	GFPGAN
RAF-DB	80.48%	76.75%	79.19%	77.52%

Table A.12: Facial Emotion classification Balanced Accuracy score.

	Balanced Accuracy ↑			
Dataset	Original	RestoreFormer	PSFRGAN	GFPGAN
RAF-DB	70.99%	64.14%	68.70%	63.79%

Table A.13: Facial Emotion Classification ΔAccuracy.

	ΔAccuracy ↑		
Dataset	RestoreFormer	PSFRGAN	GFPGAN
RAF-DB	-3.73%	-1.29%	-2.96%

Table A.14: Facial Emotion Classification ΔBalanced Accuracy.

	ΔBalanced Accuracy ↑		
Dataset	RestoreFormer	PSFRGAN	GFPGAN
RAF-DB	-6.85%	-2.29%	-7.20%

Table A.15: Identification classification Accuracy score.

	Accuracy ↑			
Dataset	Original	RestoreFormer	PSFRGAN	GFPGAN
VggFace2-Test	98.52%	97.83%	98.17%	97.91%

Table A.16: Identification classification Balanced Accuracy score.

	Balanced Accuracy ↑			
Dataset	Original	RestoreFormer	PSFRGAN	GFPGAN
VggFace2-Test	98.13%	97.30%	97.68%	97.34%

Table A.17: Identification Classification ΔAccuracy.

	ΔAccuracy ↑		
Dataset	RestoreFormer	PSFRGAN	GFPGAN
VggFace2-Test	-0.69%	-0.35%	-0.61%

Table A.18: Identification Classification ΔBalanced Accuracy.

	ΔBalanced Accuracy ↑		
Dataset	RestoreFormer	PSFRGAN	GFPGAN
VggFace2-Test	-0.83%	-0.45%	-0.79%

Table A.19: Class Rate f of Original Labels (OL), Original Predictions (OP), RestoreFormer (RF), PSFRGAN (PG), GFPGAN (GG). Values for each soft biometrics are averaged on all the datasets with available labels for that soft biometric. The average is weighted on the number of samples.

		Class Rate f				
		OL	OP	RF	PG	GG
Gender	**Male**	69.56%	69.67%	69.43%	69.26%	69.83%
	Female	30.44%	30.33%	30.57%	30.74%	30.17%
Age	**Mean**	37.03	36.82	36.04	35.55	34.89
	Variance	224.62	194.18	177.74	168.65	170.35
	Std. Deviation	14.64	13.79	13.19	12.84	12.93
Ethnicity	**CL**	66.92%	68.63%	69.97%	69.25%	69.98%
	AA	10.04%	9.16%	9.11%	8.76%	8.62%
	EA	13.12%	12.35%	12.16%	12.31%	12.64%
	AI	9.92%	9.86%	8.76%	9.68%	8.76%
F. Emotion	**Surprise**	10.35%	11.05%	11.68%	11.01%	11.40%
	Fear	2.58%	2.20%	1.67%	1.95%	1.32%
	Disgust	5.51%	3.87%	3.56%	3.66%	3.42%
	Happiness	39.77%	38.90%	39.80%	39.25%	40.05%
	Sadness	13.45%	14.05%	12.72%	14.19%	11.26%
	Anger	5.54%	6.17%	5.72%	5.89%	5.02%
	Neutral	22.80%	23.77%	24.85%	24.05%	27.54%

Table A.20: Variation is Class Rate Δf of Original Predictions (OP), RestoreFormer (RF), PSFRGAN (PG), GFPGAN (GG) with respect to Original Labels (OL). Values for each soft biometrics are averaged on all the datasets with available labels for that soft biometric. The average is weighted on the number of samples.

Class Rate Variation Δf					
		OP	RF	PG	GG
Gender	**Male**	0.11%	-0.13%	-0.29%	0.27%
	Female	-0.11%	0.13%	0.29%	-0.27%
Age	**Mean**	-0.2114	-0.9937	-1.4788	-2.1411
	Variance	-30.4422	-46.8738	-55.9707	-54.2634
	Std. Deviation	-0.8481	-1.4519	-1.7988	-1.7108
Ethnicity	**CL**	1.71%	3.05%	2.33%	3.06%
	AA	-0.88%	-0.93%	-1.28%	-1.42%
	EA	-0.77%	-0.97%	-0.81%	-0.48%
	AI	-0.06%	-1.16%	-0.24%	-1.16%
F. Emotion	**Surprise**	0.70%	1.32%	0.66%	1.05%
	Fear	-0.38%	-0.91%	-0.63%	-1.25%
	Disgust	-1.64%	-1.95%	-1.85%	-2.09%
	Happiness	-0.87%	0.03%	-0.52%	0.28%
	Sadness	0.59%	-0.73%	0.73%	-2.20%
	Anger	0.63%	0.17%	0.35%	-0.52%
	Neutral	0.98%	2.06%	1.25%	4.74%