International Journal of Computer Vision (2021) 129:2288–2307
https://doi.org/10.1007/s11263-021-01448-w

Mitigating Demographic Bias in Facial Datasets with Style-Based Multi-attribute Transfer

Markos Georgopoulos[1] · James Oldfield[2] · Mihalis A. Nicolaou[2] · Yannis Panagakis[3] · Maja Pantic[1]

Received: 15 May 2020 / Accepted: 20 February 2021 / Published online: 15 May 2021

Abstract

Deep learning has catalysed progress in tasks such as face recognition and analysis, leading to a quick integration of technological solutions in multiple layers of our society. While such systems have proven to be *accurate* by standard evaluation metrics and benchmarks, a surge of work has recently exposed the demographic bias that such algorithms exhibit–highlighting that *accuracy* does not entail *fairness*. Clearly, deploying biased systems under real-world settings can have grave consequences for affected populations. Indeed, learning methods are prone to inheriting, or even amplifying the bias present in a training set, manifested by uneven representation across demographic groups. In facial datasets, this particularly relates to attributes such as *skin tone*, *gender*, and *age*. In this work, we address the problem of mitigating bias in facial datasets by data augmentation. We propose a multi-attribute framework that can successfully transfer complex, multi-scale facial patterns *even* if these belong to underrepresented groups in the training set. This is achieved by relaxing the rigid dependence on a single attribute label, and further introducing a tensor-based mixing structure that captures multiplicative interactions between attributes in a multilinear fashion. We evaluate our method with an extensive set of qualitative and quantitative experiments on several datasets, with rigorous comparisons to state-of-the-art methods. We find that the proposed framework can successfully mitigate dataset bias, as evinced by extensive evaluations on established *diversity* metrics, while significantly improving fairness metrics such as equality of opportunity.

Keywords Data augmentation · Style transfer · Dataset bias · Demographic bias · Algorithmic fairness · Diversity · Age progression

1 Introduction

Deep learning-based models have been successfully utilized to advance the state-of-the-art in face analysis, resulting in accurate algorithms for the recognition of the identity Masi et al. (2018), age (Fu et al. 2010; Georgopoulos et al. 2018), gender, (Ng et al. 2012) and expressions (Li and Deng 2020) of the human face. Building on this success, automatic face analysis facilitates modern human-computer interaction and has found application in numerous areas of everyday life. Most importantly, machine learning and computer vision models have been applied for critical decision-making tasks,

Communicated by Daniel Kondermann.

Markos Georgopoulos and James Oldfield contributed equally.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-021-01448-w.

✉ Markos Georgopoulos
m.georgopoulos@imperial.ac.uk

James Oldfield
j.oldfield@cyi.ac.cy

Mihalis A. Nicolaou
m.nicolaou@cyi.ac.cy

Yannis Panagakis
yannisp@di.uoa.gr

Maja Pantic
m.pantic@imperial.ac.uk

[1] Department of Computing, Imperial College London, London, UK

[2] Computation-Based Science and Technology Research Center, The Cyprus Institute, Nicosia, Cyprus

[3] Department of Informatics and Telecommunications, University of Athens, Athens, Greece

from predicting recidivism and criminal behavior, to assessing candidates in interviews, and automating border control. With the wide adoption of this technology, its developers bear the responsibility to ensure that these systems do not discriminate against any subpopulation of users, i.e., that they are *fair*. However, machine learning models are prone to inheriting or even amplifying the bias that is present in the training data. In the context of face analysis, an algorithm can perform unfairly when applied on demographic groups that are underrepresented in the training set (e.g., faces of a specific gender, skin tone, or age group). This is despite the fact that the algorithm may appear as *accurate* given current evaluation metrics, but inherently fail to capture properties such as *fairness* and *diversity*.

In recent years, a surge of work has exposed the demographic bias of face analysis systems. In Buolamwini and Gebru (2018), Buolamwini and Gebru showed that commercial gender classification systems performed significantly worse on darker-skinned females. Moreover, state-of-the-art face recognition models have been reported to demonstrate bias with regards to the age, gender, and skin tone of the input face (Serna et al. 2019; Wang et al. 2019; Nagpal et al. 2019). In most cases, these demographic disparities in model performance are caused by the lack of diversity in the publicly available face datasets (Kuhlman et al. 2020; Holstein et al. 2019). For instance, widely adopted facial datasets like LFW (Huang et al. 2008) and CelebA (Liu et al. 2015) contain mainly faces of lighter skin. Similarly, only 0.5% of the faces in MOPRH (Ricanek and Tesafaye 2006) are of people over 60 years, while 87% of the faces in FG-NET (Lanitis 2002) are younger than 30 years old. Nevertheless, collecting a diverse dataset large enough for modern deep learning tasks is a herculean and tedious task. Thus, modern state-of-the-art face analysis systems are trained and tested on datasets that are either lacking in size or diversity. To circumvent this issue, practitioners turn to the plethora of available data augmentation techniques [see Shorten and Khoshgoftaar (2019) for a survey].

Data augmentation methods range from simple image transformations (e.g., mirroring, rotation, and random cropping) to non-photorealistic image mixing [e.g., Inoue (2018), Zhang et al. (2017)], and deep generative models Sandfort et al. (2019). In the latter category, neural style transfer has proved to be an efficient data augmentation tool, that can be used to train robust classifiers Jackson et al. (2019), (Perez and Wang 2017; Zheng et al. 2019). In this spirit, we propose a novel style transfer framework that is tailored to the task of diversity-enhancing data augmentation. Since our aim is to enhance the demographic diversity of a facial dataset, we propose a style transfer approach using Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) that is able to transfer multiple demographic attributes for each image in a biased set. The resulting image set is less ridden with demographic biases, and can hence be used to train fairer face classifiers.

Translation of demographic attributes has been studied extensively, albeit for individual attributes. In particular, facial *aging* is a long-standing task in computer vision (Ramanathan et al. 2009; Fu et al. 2010; Georgopoulos et al. 2018) with recent GAN-based approaches being able to produce realistically aged and rejuvenated faces (Zhang et al. 2017; Wang et al. 2018; Duong et al. 2019). On the other hand, modifying the gender of a face is a common application of facial attribute transfer (Choi et al. 2018; He et al. 2019). However, these methods are not capable of synthesizing realistic samples at the tails of the distribution—e.g., for faces over 60 years old—as we show experimentally in Sect. 4.4. At the same time, face synthesis in such models are usually conditioned on a single attribute label. Hence, these methods are only able to generate a single image per attribute class, that is constrained by the bias of the training set. The aforementioned limitations prevent these approaches from being able to efficiently mitigate dataset bias.

In this work, instead of collapsing attribute information into a single label, we condition the face generation on *discriminative representations* for each attribute. For instance, despite training with the provided binary gender labels, our method learns a high-dimensional, continuous-valued representation of gender, which better reflects the underlying non-binary nature of the attribute. Motivated by the recent success of style-based GANs (Karras et al. 2019; Huang et al. 2018; Kim et al. 2020; Park et al. 2019; Ma et al. 2018) in transferring arbitrary styles at different scales, we propose to transfer the joint demographic style of each subpopulation. Consequently, our method is able to modify the attributes of each face and enhance the diversity of the dataset. In order to combine the different representations capturing complex facial patterns related to each attribute, we propose a novel extension to AdaIN (Huang and Belongie 2017). The proposed method is tailored to handle the mixing of multiple attribute representations by introducing a tensor-based mixing structure that captures multiplicative attribute interactions in a multilinear fashion, effectively facilitating multi-attribute transfer given a single input image.

As we show in this paper, the proposed formulation is flexible enough to synthesize images of large intra-class diversity, transferring complex facial patterns *even* for samples that belong to the tails of the training set distribution. Summarizing, the contributions of this paper are listed in what follows.

– We introduce a novel style transfer GAN that is able to transfer multiple demographic attributes simultaneously. By conditioning on different sets of target images, our framework is able to generate diverse images for each attribute class.

– A multi-attribute extension to AdaIN is presented in Sect. 3.2. The proposed fusion framework is able to model the multiplicative interactions between attribute representations by employing a tensor-based mixing structure, resulting into a single conditioning variable.
– With a series of qualitative and quantitative experiments (Sect. 4), we benchmark our model's ability to enhance the diversity of a dataset against state-of-the-art baselines (both multi-attribute transfer and age progression methods). To quantify the diversity enhancing capabilities of the models, we turn to the established diversity metrics introduced in Merler et al. (2019).
– We provide a thorough investigation of how dataset bias affects classification performance, and show how more diverse datasets can be used to train less biased classifiers (Sect. 4.6). We also study the case of bias in gender recognition—within the binary paradigm in which it is currently commonly framed in practice—on two datasets: MORPH and KANFace. The experimental analysis indicates that by augmenting the training sets using our model, we are able to mitigate the classifier biases more effectively than other, state-of-the-art methods.

The proposed framework is based on our prior work (Georgopoulos et al. 2020) that used standard generative architectures that could be adapted to perform face aging, focusing solely on age progression. This work significantly extends our preliminary work as it is designed to transfer multiple attributes instead of just age, by proposing a novel multilinear extension to AdaIN that is suitable for multiple attributes. That is, the proposed method is able to handle the mixing of multiple attribute representations in a multi-linear fashion. The style transfer model is evaluated on additional datasets and compared to multi-attribute GANs [i.e., StarGAN (Choi et al. 2018) and AttGAN (He et al. 2019)]. Lastly, the focus of the preliminary work was dataset diversity. In this work, we make the connection between diversity and algorithmic fairness. In particular, we enhance the diversity of datasets and evaluate the fairness of the classifiers that are trained on them, offering a thorough comparison to 7 state-of-the-art bias mitigation methods on 2 datasets.

2 Related Work

In this section, we provide an overview of related work from the area of style-based image transfer with adversarial learning (Sect. 2.1). Furthermore, we introduce related work in terms of demographic attribute editing (Sect. 2.2), as several methods have been proposed in literature for editing such attributes; albeit with most works focusing on one attribute individually. Finally, we discuss related literature on fairness and bias mitigation, focusing on face analysis (Sect. 2.3).

2.1 Generative Adversarial Networks and Style Transfer

Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) are the driving force behind the recent success of deep generative models in image synthesis (Radford et al. 2015; Karras et al. 2019; Brock et al. 2018; Karras et al. 2017; Park et al. 2019) and image-to-image translation (Choi et al. 2018; Huang et al. 2018; Kim et al. 2020; Isola et al. 2017; Zhu et al. 2017; Tang et al. 2019). Multiple variations have been proposed, including different training objectives (Salimans et al. 2016; Arjovsky et al. 2017; Mao et al. 2017; Lim and Ye 2017) and architectural choices (Radford et al. 2015; Karras et al. 2019). Among these variations, style-based GANs draw inspiration from the style transfer literature (Gatys et al. 2015) and use Adaptive Instance Normalization (AdaIN) (Huang and Belongie 2017) to condition image generation. MUNIT (Huang et al. 2018) uses AdaIN in a GAN setting to perform image-to-image translation by injecting style content into an autoencoder-like network at the bottleneck layers. The seminal work of StyleGAN (Karras et al. 2019) proposed a generator architecture using AdaIN to modulate style content at multiple resolutions. The state-of-the-art image-to-image translation methods continue to adopt style-based approaches (Huang et al. 2018; Kim et al. 2020; Park et al. 2019; Ma et al. 2018), and we follow in this vein–treating demographic attributes as target styles.

2.2 Transfer of Demographic Attributes

Synthesizing faces of a specific target demographic has been studied extensively, albeit for individual demographic attributes. In particular, age progression refers to the task of rendering an aged or rejuvenated image of an input face (Fu et al. 2010; Ramanathan et al. 2009; Georgopoulos et al. 2018). While earlier works in age progression proposed simplistic prototype-based approaches that could not produce photorealistic results, a number of recently proposed GAN-based methods are capable of convincing face aging. Zhang et al. (2017) introduced a conditional adversarial autoencoder (CAAE) that models aging as the traversal of a low-dimensional manifold. GAN-based image translation approaches were also proposed in Wang et al. (2018), Yang et al. (2018), where pre-trained networks were used to facilitate identity preservation and aging accuracy, respectively. Besides age, generating faces with different genders is a standard benchmark of the CELEB-A dataset (Liu et al. 2015), and is successfully performed by numerous attribute editing approaches (Choi et al. 2018; He et al. 2019, 2017; Perarnau et al. 2016; Li et al. 2016). Lastly, in order to mitigate racial

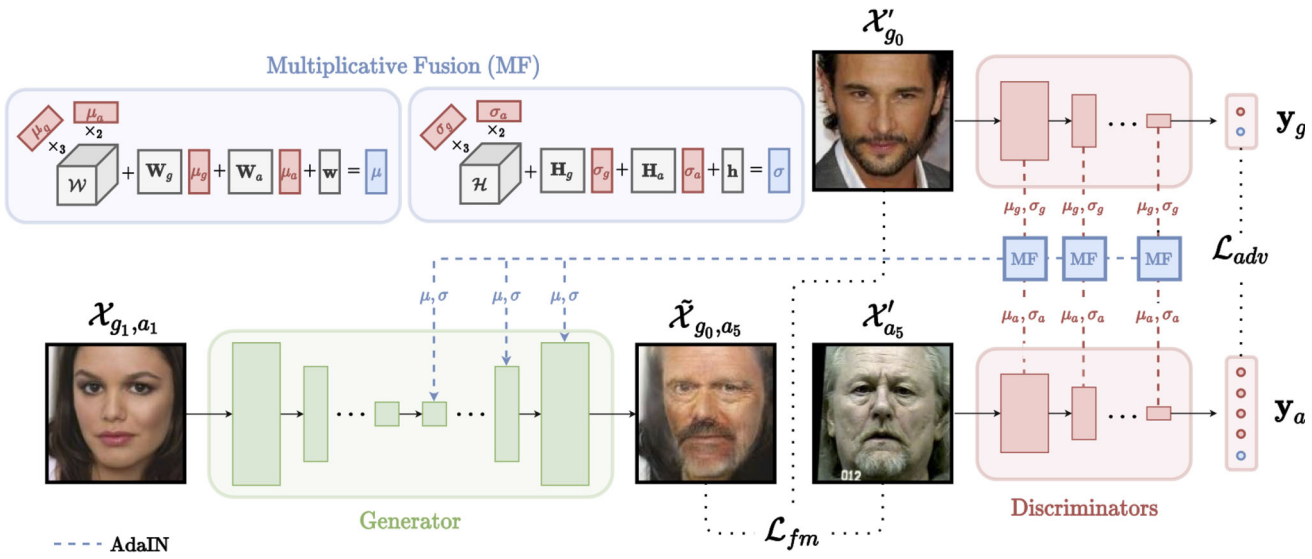


Fig. 1 Overview of the proposed method for multi-attribute transfer by way of example: An input image $\mathcal{X}_{g_1,a_1}$ of gender g_1 and age a_1 is passed through the Generator (left), to translate it to target age a_5 and gender g_0. At each upsampling block in the generator, we perform AdaIN to modulate the style content. The new statistics are computed using a multiplicative fusion module that captures the interactions between the discriminators' activations' moments when evaluated on target images of the desired classes. The logits output from the discriminators are used to train our adversarial loss $\mathcal{L}_{adv}$, and the target and synthetic images are used to train the generator's $\mathcal{L}_{fm}$ feature-matching loss

bias in face recognition, Yucer et al. (2020) proposed to use CycleGAN (Zhu et al. 2017) to transfer race. In our work, instead of focusing on a single attribute, we propose to simultaneously edit multiple demographic attributes. Unlike many stand-alone age progression methods (Wang et al. 2018; Yang et al. 2018) and style transfer-based GANs (Karras et al. 2019; Huang et al. 2018), we require no additional networks to obtain the conditioning information.

2.3 Fairness-aware Learning and Face Analysis

In light of the growing concerns regarding algorithmic discrimination, the field of fairness-aware machine learning has attracted interest from the research community. Most of the work focuses on tabular data [e.g. the UCI Adult income dataset Dua and Graff (2017)] and tackles the problem of fair classification with regards to a protected attribute [e.g., Edwards and Storkey (2015), Madras et al. (2018)]. In such cases, both the output variable and the protected attribute are binary. For instance, a common scenario in the UCI Adult dataset is the classification of the subjects into two categories based on their wage, while being fair with regards to gender. These works have given rise to different definitions of fairness (Mehrabi et al. 2019; Verma and Rubin 2018), the most common of which are: (i) demographic parity, (ii) equalized odds, and (iii) equality of opportunity. Demographic parity ensures that the predicted label $\hat{Y}$ is independent of the protected attribute S, i.e., $P(\hat{Y} = 1|S = 1) = P(\hat{Y} = 1|S = 0)$. Unlike demographic parity, a predictor $\hat{Y}$ that satisfies equalized odds can depend on S, but only through the target variable Y (Hardt et al. 2016), that is: $P(\hat{Y} = 1|S = 1, Y = y) = P(\hat{Y} = 1|S = 0, Y = y), y \in \{0, 1\}$. This definition implies that both the true and false positive rates will be the same for each population. Equality of opportunity is a relaxed notion of equalized odds, that only requires the true positive rates to be equal (Hardt et al. 2016), formally: $P(\hat{Y} = 1|S = 1, Y = 1) = P(\hat{Y} = 1|S = 0, Y = 1)$. In this work, we use equality of opportunity to quantitatively evaluate the fairness of our face analysis models as it one of the most commonly employed approaches.

It is only very recently that fairness-aware algorithms for face analysis have attracted similar attention. The experimental results in Buolamwini and Gebru (2018) indicate that commercial gender recognition systems demonstrate a significant difference in performance between lighter male and darker female faces. This study resulted in replies from the vendors that focused on the importance of diversity in the training datasets (Raji and Buolamwini 2019). Thereafter, a series of in-processing and pre-processing methods have been proposed for fair face analysis–the former targeting unfairness at the algorithm-level and the latter at the data-level. In the former category, Alvi et al. (2018) proposed a joint learning and un-learning framework, while Kim et al. (2019) learn a fair classifier by minimizing the mutual information between the intermediate representation and the bias. These methods employ techniques from domain adaptation to learn a representation that minimizes classification loss while being invariant to the sensitive attribute. In the lat-

ter category, Sattigeri et al. (2018) extend AC-GAN (Odena et al. 2017) to generate a fair dataset, while Quadrianto et al. (2018) use an autoencoder to remove sensitive information from images. In this work, we introduce an image-to-image translation model to augment the training set, and thus, our framework is most closely related to Quadrianto et al. (2018). However, contrary to Quadrianto et al. (2018), our neural style transfer approach is able to generate naturalistic faces by translating demographic attributes instead of removing them (e.g., gender-less faces).

3 Methodology

In this section, we introduce the proposed multi-attribute style transfer framework that enhances the diversity of a face dataset suffering from demographic bias. Our aim is to utilize the proposed generative model to augment a biased training set. Thus, by training a face analysis model on the augmented set, we can mitigate the demographic bias and achieve fairer classification performance.

Drawing inspiration from style transfer GAN literature, the proposed model is able to transfer the joint demographic attribute style from a set of images. In particular, we utilize discriminative features for each attribute at different scales to guide the image translation using AdaIN. To obtain the joint demographic attribute style, we present a novel extension to the AdaIN framework suitable for mixing the statistics of multiple attribute representations. The proposed approach models multiplicative interactions between the attributes, leading to a single fused variable for conditioning at each generator layer.

The remainder of this section is structured as follows. In Sect. 3.1 we introduce the notation as well as the basic matrix and tensor operations used in the paper. The multi-attribute style transfer framework is presented in Sect. 3.2. The different components of the training objective are analyzed in Sect. 3.3. Finally, an overview of the proposed method is visualized in Fig. 1.

3.1 Notation

In this section, we introduce the notation and definitions of the operations used throughout this paper. Concretely, we denote tensors by calligraphic letters, e.g., $\mathcal{X}$, matrices by uppercase boldface letters, e.g., $\mathbf{X}$ and vectors by lowercase boldface letters, e.g., $\mathbf{x}$. We refer to each element of a K^{th} order tensor $\mathcal{X}$ by K indices, i.e., $(\mathcal{X})_{i_1 i_2 \ldots i_K} \doteq x_{i_1 i_2 \ldots i_K}$. Similarly, we refer the elements of matrix $\mathbf{X}$ as x_{ij}, while $\mathbf{x}_j$ denotes the j-th column of the matrix. The D-dimensional vector of ones is denoted as $\mathbf{1}_D$.

Hadamard product The *Hadamard product of* $\mathbf{A} \in \mathbb{R}^{I \times N}$ and $\mathbf{B} \in \mathbb{R}^{I \times N}$ is denoted by $\mathbf{A} * \mathbf{B}$ and is the element-wise multiplication of the two matrices.

Kronecker product The *Kronecker product* of two matrices $\mathbf{A} \in \mathbb{R}^{I \times J}$ and $\mathbf{B} \in \mathbb{R}^{K \times L}$ is denoted by $\mathbf{A} \otimes \mathbf{B} \in \mathbb{R}^{(IK)\times(JL)}$ and is defined as:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & a_{12}\mathbf{B} & \ldots & a_{1J}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} & \ldots & a_{2J}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}\mathbf{B} & a_{I2}\mathbf{B} & \ldots & a_{IJ}\mathbf{B} \end{bmatrix}. \tag{1}$$

Khatri-Rao product The *Khatri-Rao product* of two matrices $\mathbf{A} \in \mathbb{R}^{I \times N}$ and $\mathbf{B} \in \mathbb{R}^{J \times N}$ is denoted by $\mathbf{A} \odot \mathbf{B}$. The resulting matrix is of dimensions $(IJ) \times N$, and is defined as:

$$\mathbf{A} \odot \mathbf{B} = \begin{bmatrix} \mathbf{a}_1 \otimes \mathbf{b}_1 & \mathbf{a}_2 \otimes \mathbf{b}_2 & \cdots & \mathbf{a}_N \otimes \mathbf{b}_N \end{bmatrix}. \tag{2}$$

Tensor unfolding The *mode-k unfolding* of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \ldots \times I_K}$ reorders the elements of $\mathcal{X}$ into a matrix $\mathbf{X}_{(k)} \in \mathbb{R}^{I_k \times \bar{I}_k}$ with $\bar{I}_k = \prod_{\substack{t=1 \\ t \neq k}}^{K} I_t$.

Mode-k vector product The *mode-k vector product* of a tensor $\mathcal{X} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_K}$ and a vector $\mathbf{v} \in \mathbb{R}^{I_k}$ is denoted by $\mathcal{X} \times_k \mathbf{v} \in \mathbb{R}^{I_1 \times \ldots \times I_{k-1} \times I_{k+1} \times \ldots \times I_K}$, defined elementwise as:

$$(\mathcal{X} \times_k \mathbf{v})_{i_1 \ldots i_{k-1} i_{k+1} \ldots i_K} = \sum_{i_k=1}^{I_k} x_{i_1 i_2 \ldots i_K} v_{i_k}. \tag{3}$$

CP decomposition The CANDECOMP/PARAFAC (CP) tensor decomposition (Carroll and Chang 1970; Harshman 1970) factorizes a tensor into a sum of rank-one tensors. Let $\mathcal{X}$ be a tensor of rank R. Its CP decomposition is:

$$\mathcal{X} \doteq [\![\mathbf{U}_{[1]}, \mathbf{U}_{[2]}, \ldots, \mathbf{U}_{[K]}]\!] = \sum_{r=1}^{R} \mathbf{u}_r^{(1)} \circ \mathbf{u}_r^{(2)} \circ \cdots \circ \mathbf{u}_r^{(K)}, \tag{4}$$

where $\circ$ denotes the vector outer product. The matrices $\left\{\mathbf{U}_{[k]} = [\mathbf{u}_1^{(k)}, \mathbf{u}_2^{(k)}, \cdots, \mathbf{u}_R^{(k)}] \in \mathbb{R}^{I_k \times R}\right\}_{k=1}^{K}$ consist of the rank-one components. The CP decomposition of a third order tensor $\mathcal{X}$ is written in matrix form as Kolda and Bader (2009):

$$\mathbf{X}_{(1)} \doteq \mathbf{U}_{[1]} \left(\mathbf{U}_{[3]} \odot \mathbf{U}_{[2]} \right)^T. \tag{5}$$

3.2 Proposed Framework

Style-conditioned Generator For simplicity of presentation, we consider the case of two attributes, namely gender and age. The generator of our framework is an autoencoder G that learns a mapping from a face $\mathcal{X}_{g,a}$ of gender g and age

a to a synthesized face $\tilde{\mathcal{X}}_{g',a'}$ of gender g' and age a'. To achieve this, G uses the gender features of a target image $\mathcal{X}'_{g'}$ and the age features of target image $\mathcal{X}'_{a'}$, which are extracted using the corresponding discriminators D^{gen} and D^{age}. We consider these features to capture the *demographic* styles $\mathbf{s}_g$ and $\mathbf{s}_a$. The styles are obtained from the target images as the activations at different layers of the discriminator networks. The age and gender styles are then fused and injected into the decoder of the generator using the proposed multi-attribute AdaIN, which is described in detail below.

Multi-attribute AdaIN The standard approach to style transfer is conditioning the generation on a single style representation. This is achieved by using AdaIN at different layers of the generator. The AdaIN operator scales and shifts the normalized activations at each layer of the decoder, in order to match a target style. We denote $\mathcal{Z}^i \in \mathbb{R}^{d_h^i \times d_w^i \times d_c^i}$ the activation of the i-th layer of the generator and its corresponding target style s^i which is represented by a set of two vectors, i.e., $s^i \doteq \{\boldsymbol{\mu}^i, \boldsymbol{\sigma}^i\}$. In this work, the vectors $\boldsymbol{\mu}^i$ and $\boldsymbol{\sigma}^i$ are obtained from a target image as the channel-wise first and second order moments of the activations at each layer of the discriminators. We cast vectors $\boldsymbol{\mu}^i$ and $\boldsymbol{\sigma}^i$ to have the same dimensions as $\mathcal{Z}$:

$$\mathcal{M}^i = \mathbf{1}_{d_h^i} \circ \mathbf{1}_{d_w^i} \circ \boldsymbol{\mu}^i, \tag{6}$$

$$\mathcal{S}^i = \mathbf{1}_{d_h^i} \circ \mathbf{1}_{d_w^i} \circ \boldsymbol{\sigma}^i, \tag{7}$$

where $\mathcal{M}^i, \mathcal{S}^i \in \mathbb{R}^{d_h^i \times d_w^i \times d_c^i}$. Then, the AdaIN operator is defined as:

$$\text{AdaIN}(\mathcal{Z}^i, s^i) = \mathcal{M}^i + \frac{\mathcal{Z}^i - \mu(\mathcal{Z}^i)}{\sigma(\mathcal{Z}^i)}\mathcal{S}^i, \tag{8}$$

where $\mu(\mathcal{Z}^i)$ and $\sigma(\mathcal{Z}^i)$ denote the tensorized channel-wise mean and variance of the activations $\mathcal{Z}^i$. In this work, we aim to transfer a collection of styles (i.e., attributes) that are captured by different activations. To this end, we introduce a multi-linear style mixing model that models the multiplicative interactions (Jayakumar et al. 2020) between the activations for each attribute, by assuming a tensorial mixing structure. Concretely, given target styles for age $s_a^i \doteq \{\boldsymbol{\mu}_a^i, \boldsymbol{\sigma}_a^i\}$ and gender $s_g^i \doteq \{\boldsymbol{\mu}_g^i, \boldsymbol{\sigma}_g^i\}$ at layer i of the discriminators, we propose the following factorization:

$$\boldsymbol{\mu}^i = \mathbf{w}^i + \mathbf{W}_a^i \boldsymbol{\mu}_a^i + \mathbf{W}_g^i \boldsymbol{\mu}_g^i + \mathcal{W}^i \times_2 \boldsymbol{\mu}_a^i \times_3 \boldsymbol{\mu}_g^i, \tag{9}$$

which can be written (as shown in Kolda (2006), Kolda and Bader (2009)) as:

$$\boldsymbol{\mu}^i = \mathbf{w}^i + \mathbf{W}_a^i \boldsymbol{\mu}_a^i + \mathbf{W}_g^i \boldsymbol{\mu}_g^i + \mathbf{W}_{(1)}^i (\boldsymbol{\mu}_g^i \odot \boldsymbol{\mu}_a^i), \tag{10}$$

where $\mathbf{W}_{(1)}^i$ is the mode-1 unfolding of tensor $\mathcal{W}^i$. The fused second order statistics are calculated from $\boldsymbol{\sigma}_a^i$ and $\boldsymbol{\sigma}_g^i$ in a similar fashion:

$$\boldsymbol{\sigma}^i = \mathbf{h}^i + \mathbf{H}_a^i \boldsymbol{\sigma}_a^i + \mathbf{H}_g^i \boldsymbol{\sigma}_g^i + \mathbf{H}_{(1)}^i (\boldsymbol{\sigma}_g^i \odot \boldsymbol{\sigma}_a^i). \tag{11}$$

The multiplicative interactions between the features are captured by tensors $\{\mathcal{W}^i, \mathcal{H}^i\} \in \mathbb{R}^{d_c^i \times d_c^i \times d_c^i}$. Due to the dimensionality of the higher order tensors ($d_c^i 3$ parameters) the number of parameters of the fusion model scales exponentially. Indicatively, a convolutional layer of the discriminator with 256 filters would require a tensor of $16M$ parameters to fuse the activations (statistics) of two attributes. To avoid this, tensors $\mathcal{W}^i$ and $\mathcal{H}^i$ assume a CP decomposition, and Eqs. (10) and (11) become:

$$\boldsymbol{\mu}^i = \mathbf{w}^i + \mathbf{W}_a^i \boldsymbol{\mu}_a^i + \mathbf{W}_g^i \boldsymbol{\mu}_g^i + \mathbf{U}_{[1]}^i (\mathbf{U}_{[3]}^i \odot \mathbf{U}_{[2]}^i)^T (\boldsymbol{\mu}_g^i \odot \boldsymbol{\mu}_a^i), \tag{12}$$

$$\boldsymbol{\sigma}^i = \mathbf{h}^i + \mathbf{H}_a^i \boldsymbol{\sigma}_a^i + \mathbf{H}_g^i \boldsymbol{\sigma}_g^i + \mathbf{V}_{[1]}^i (\mathbf{V}_{[3]}^i \odot \mathbf{V}_{[2]}^i)^T (\boldsymbol{\sigma}_g^i \odot \boldsymbol{\sigma}_a^i), \tag{13}$$

where matrices $\mathbf{U}_{[j]}^i, \mathbf{V}_{[j]}^i$ for $j \in \{1 \dots 3\}$ consist of the rank-1 tensor of the CP decompositions of tensors $\mathcal{W}^i$ and $\mathcal{H}^i$. The resulting $\boldsymbol{\mu}^i$ and $\boldsymbol{\sigma}^i$ are then used in Eqs. (6) and (8) to modulate the style in the layers of the generator.

Discriminators In the proposed framework we utilize multiple discriminator networks–one for each attribute. Similarly to Liu et al. (2019), each of the discriminators are trained to distinguish between real and fake images of each class. Since each discriminator is responsible for one attribute, the resulting representations are also discriminative with respect to that attribute. These representations are used to condition the style transfer in the decoder of the generator at each layer. Hence, we design D^{gen} and D^{age} as mirrored decoders of the generator.

3.3 Training Objective

Our full objective function comprises of two terms, namely the adversarial and the reconstruction loss. The losses are described as follows.

Adversarial loss To train the generator to synthesize photorealistic images with characteristics of the desired class, we use an adversarial loss. Given an input image $\mathcal{X}_{g,a}$ and its translation $G(\mathcal{X}_{g,a}, \mathbf{s}_{g'}, \mathbf{s}_{a'})$ to age a' and gender class g', conditioned on demographic styles $s_{g'}$ and $s_{a'}$ (obtained for each layer from $D^{gen}(\mathcal{X}'_{g'})$ and $D^{age}(\mathcal{X}'_{a'})$ respectively), we compute an adversarial loss term for each attribute y as:

$$\mathcal{L}_{adv}^y = \mathbb{E}_{\mathcal{X}_{g,a}}\big[\log D^y(\mathcal{X}_{g,a})\big] + \mathbb{E}_{\mathcal{X}_{g,a}, \mathcal{X}'_{g'}, \mathcal{X}'_{a'}}\big[\log(1 - D^y\big(G(\mathcal{X}_{g,a}, \mathbf{s}_{g'}, \mathbf{s}_{a'})\big)\big], \tag{14}$$

where D^y is the discriminator for attribute y. The combined adversarial loss for both age and gender is:

$$\mathcal{L}_{adv} = \mathcal{L}_{adv}^{gen} + \mathcal{L}_{adv}^{age} \tag{15}$$

Rather than having the generator minimize $\mathcal{L}_{adv}$, we instead adopt a feature-matching loss (Salimans et al. 2016) in order to train the generator to match the attribute patterns of particular target images. The feature-matching loss is defined as:

$$\begin{aligned}&\mathcal{L}_{fm} = \\ &\mathbb{E}_{\mathcal{X}_{g,a},\mathcal{X}'_{g},\mathcal{X}'_{a'}}\big[\, \| \, D^{gen}(\mathcal{X}'_{g'}) - D^{gen}\big(G(\mathcal{X}_{g,a}, \mathbf{s}_{g'}, \mathbf{s}_{a'})\big) \, \|_2^2 \\ &+ \| \, D^{age}(\mathcal{X}'_{a}) - D^{age}\big(G(\mathcal{X}_{g,a}, \mathbf{s}_{g'}, \mathbf{s}_{a'})\big) \, \|_2^2 \,\big]\end{aligned} \tag{16}$$

Reconstruction loss Our framework is trained to edit the selected demographic attributes, while preserving the remaining input information. To this end, we use a cycle constraint (Zhu et al. 2017) to ensure that the synthetic images can be translated back to the original input:

$$\begin{aligned}&\mathcal{L}_{rec} = \\ &\mathbb{E}_{\mathcal{X}_{g,a},\mathcal{X}_{g'},\mathcal{X}_{a'}}\big[\, \| \, \mathcal{X}_{g,a} - G\big(G(\mathcal{X}_{g,a}, \mathbf{s}_{g'}, \mathbf{s}_{a'}), \mathbf{s}_{g}, \mathbf{s}_{a}\big) \, \|_1 \,\big],\end{aligned} \tag{17}$$

where a, g are the input labels for $\mathcal{X}$'s age and gender respectively.

Full objective The full objective functions for D and G are:

$$\mathcal{L}_D = -\mathcal{L}_{adv} \tag{18}$$

$$\mathcal{L}_G = \mathcal{L}_{fm} + \lambda_{rec}\mathcal{L}_{rec}, \tag{19}$$

where λ_{rec} is the hyper-parameters for the reconstruction loss terms.

4 Experiments

In this section, we present a series of experiments designed to evaluate the efficacy of the proposed style-based model and method for mitigating dataset bias via data augmentation. Firstly, we outline in Sect. 4.1 the implementation details for all the experiments that follow. We then provide information in Sect. 4.2 about the datasets on which we evaluate the method, along with our choice of baselines in Sect. 4.3. Subsequently, in Sect. 4.4, we demonstrate the ability of our method to synthesize realistic facial images covering all demographic groups, particularly the ones at the tails of the training set distribution, e.g., over 60 years old. The diversity enhancing capabilities of the model are then evaluated in Sect. 4.5. In particular, we use the proposed method to augment an existing facial image dataset, while subsequently quantifying the diversity of the augmented set using established metrics such as the Shannon and Simpson indices introduced in Merler et al. (2019). Finally, we investigate the impact of diversity on the performance of face analysis models in Sect. 4.6. Focusing on gender recognition, we showcase that the classifiers demonstrate various demographic biases that can be successfully mitigated using the proposed neural augmentation method.

In this work, we benchmark our model on widely adopted datasets [e.g., Zhang et al. (2017), Ricanek and Tesafaye (2006)] that are annotated with binary gender labels. In particular, the datasets are annotated with the sex labels 'Male' and 'Female', and therefore it is common practice in both the face analysis (Ng et al. 2012; Dantcheva et al. 2015) and fairness literature (Buolamwini and Gebru 2018; Quadrianto et al. 2019; Zhao et al. 2017; Hendricks et al. 2018) to use these available labels under this categorization. As such, we can only address gender bias within this imposed binary classification paradigm. We note however that gender is widely considered to be non-binary,[1] and as such, inappropriate categorization of this attribute runs the risk of resulting in unfair face analysis systems. Furthermore, similar to Buolamwini and Gebru (2018) we investigate bias with regards to "skin tone". That is, using the "race" labels of MORPH, we classify the faces into "light-skinned" or "dark-skinned".

4.1 Implementation Details

We adopt the same generator architecture as StarGAN (Choi et al. 2018), with 6 layers in the encoder-decoder and 6 residual blocks (He et al. 2016) in the bottleneck. We depart from their choice of architecture for our discriminators however, and use a simple 3 layered CNN to match the dimensionality of the decoder (more details on the model architectures can be found in the supplementary material). Following recent progress in stabilising the training of GANs (Mescheder et al. 2018), we use the R_1 gradient penalty loss term in the discriminator's objective:

$$\mathcal{L}_{gp} = \lambda_{gp}\mathbb{E}_{\mathcal{X}}\Big[\, \| \, \nabla\mathcal{D}(\mathcal{X}) \, \|^2\Big]. \tag{20}$$

where $\mathcal{X}$ denotes the real images sampled from the true data distribution.

To further encourage the translated images to contain class-relevant attribute patterns, we find it beneficial to use the style parameters extracted from the *synthetic* images as the target styles in the reconstruction term in Eq. (19). Concretely, rather than translating the synthetic images back to the input images using the style parameters s_g, s_a from the

[1] https://www.ons.gov.uk/economy/environmentalaccounts/articles/whatisthedifferencebetweensexandgender/2019-02-21.

input images, we instead use $\tilde{s}_g$ from $D^{gen}(\tilde{\mathcal{X}}_{g,a})$ and $\tilde{s}_a$ from $D^{age}(\tilde{\mathcal{X}}_{g,a})$. By requiring the synthetic images to retain sufficient style information to reconstruct the original images, the reconstruction term thus also aids in the attribute transfer.

We train our networks with the hyperparameters outlined in Liu et al. (2019): $\lambda_{rec} = 0.01$, $\lambda_{gp} = 10.0$, and opt to train our networks end-to-end with Kingma and Ba (2014), with a learning rate of 10^{-4}, and $\beta_1 = 0.5$, $\beta_2 = 0.99$. For the multiplicative fusion layers, we set the rank of the tensors in the CP decomposition to be equal to half the number of filters at each layer.

4.2 Datasets

We conduct experiments on a number of popular databases with demographic attribute annotations. Concretely, we adopt the **MORPH** (Ricanek and Tesafaye 2006), **CACD** (Chen et al. 2014), **KANFace** (Georgopoulos et al. 2020), and **UTKFace** (Zhang et al. 2017) datasets. MORPH's second album features 55,134 near-frontal facial images of 13,618 identities–captured in a controlled setting and demonstrate little variation in expressions, illumination, or background. For our experiments on bias mitigation, we consider all three of the given annotated attributes: *'age'*, *'gender'*, and *'race'*. We use images with the attribute *skin tone* labelled as either dark-skinned or light-skinned, which amounts to 53,140 facial images, covering more than 96% of the dataset. In contrast, the CACD dataset features images captured in-the-wild, collected from Google Images. It contains over 160,000 images from 2000 celebrities. The dataset is annotated with regards to age. We manually annotate gender, using the provided identities. Given these two annotated attributes, we model both *age* and *gender* for our experiments on CACD. In Sect. 4.6, we train classifiers in the augmented MORPH and KANFace datasets. KANFace is the largest manually annotated video dataset of faces captured in-the-wild. We utilize the static version of KANFace that consists of 40 K images that are annotated with regards to 'identity', 'age', 'gender', and 'kinship'. Similarly to CACD, we utilize only the age and gender labels. Finally, the UTKFace dataset consists of over 20,000 images of individuals aged between 0 and 116 years old. The dataset is labelled with 'age', 'gender', and 'ethnicity' labels. Similar to MORPH, we utilize 'skin tone' labels instead of 'ethnicity'. For all datasets, we use 80% of the images for training and keep 20% for testing. The faces are grouped into 5 distinct classes: 0–18, 19–30, 31–45, 46–60, 61+. This age split is more fine-grained than the one commonly used in face aging literature (Yang et al. 2018; Wang et al. 2016; Yang et al. 2016) (i.e., 0–30, 31–40, 41–50, 51+), utilized to uncover the biases against faces under 18 and over 60 years old.

4.3 Baselines

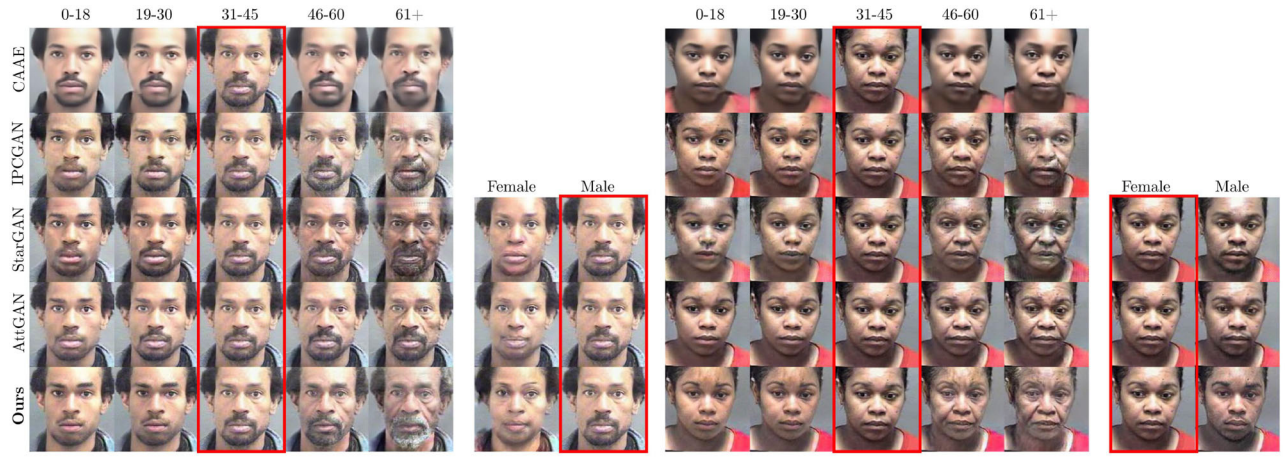
We benchmark our model against two strong baselines for multi-attribute image translation, namely **StarGAN** (Choi et al. 2018) and **AttGAN** (He et al. 2019). Additionally, since face aging has been in itself investigated in the literature, we compare against two standalone face aging GANs, namely **IPCGAN** (Wang et al. 2018) and **CAAE** (Zhang et al. 2017). Both StarGAN and AttGAN translate an input image to multiple target domains by conditioning on a one-hot encoded label vector. StarGAN differs from AttGAN in that it employs a cycle-constraint reconstruction loss, while AttGAN utilizes a so-called attribute classification constraint to encourage correct attribute classification in synthesized images. On the other hand, IPCGAN uses a separate pre-trained age classifier to enforce the aging features of a particular target age class on the translated faces, while CAAE learns face aging by traversing a manifold. We benchmark our model against these 4 baselines both qualitatively and quantitatively in the sections that follow. The baselines were implemented using the authors' publicly released code where available.[2]

4.4 Qualitative Results

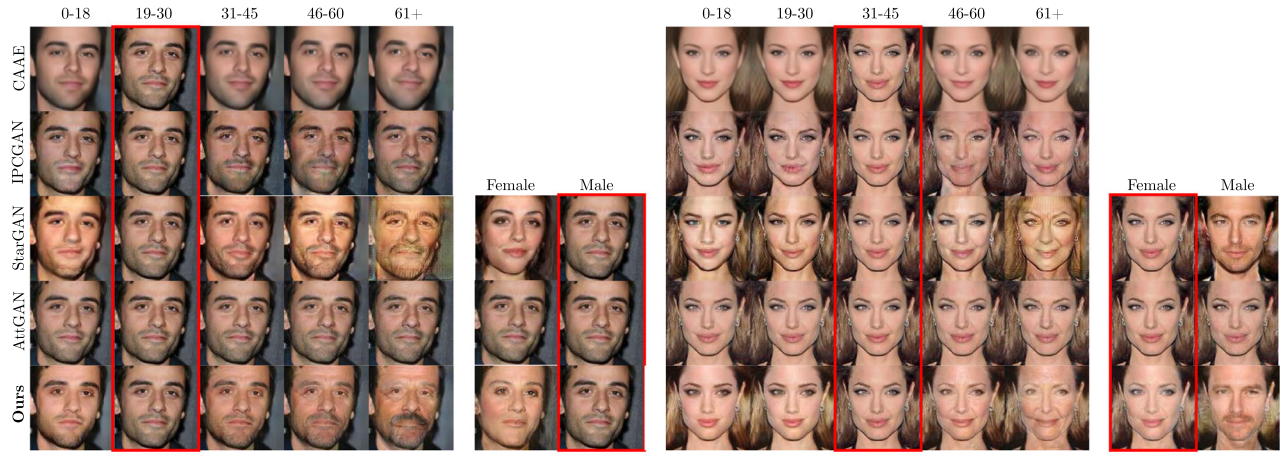
4.4.1 Attribute Transfer

Our approach to dataset bias mitigation involves augmenting the biased datasets in such way that the joint distribution of the attributes' labels is uniform. For this reason, the ability to synthesize sharp and realistic images that are well-classified as the target class is paramount. In Fig. 2a to 2c we translate a given facial image to all attribute combinations for the available class labels and compare the results of our method to the baselines. We find that both IPCGAN and CAAE fail to produce sharp images representative of the target class, *especially* for faces under 18 and over 60, due to the underrepresentation of these classes in the training set. At the same time AttGAN fails to modify the input image in a noticeable manner, especially in the case of CACD, where the generated faces are very similar to the input (e.g. row 4 of Fig. 2b). Furthermore, whilst StarGAN produces prominent class-discriminative features in the synthetic images, it fails to retain photo-realism in the generated facial images. As a result, the StarGAN's image translation results often contain artifacts. This is particularly pronounced in StarGAN's transfers to underrepresented demographic groups

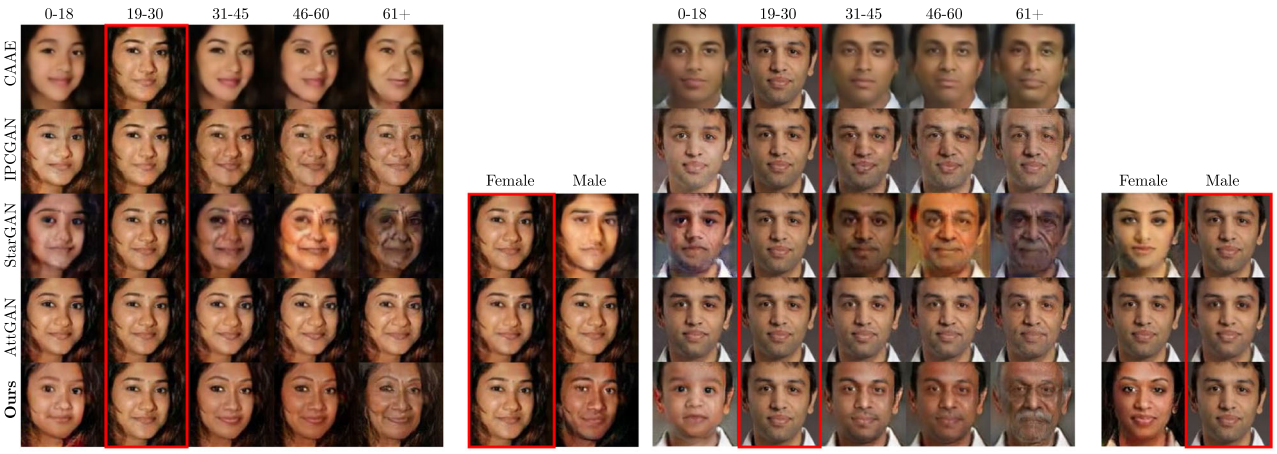
[2] **CAAE**: https://github.com/ZZUTK/Face-Aging-CAAE and, **IPCGAN**: https://github.com/dawei6875797/Face-Aging-with-Identity-Preserved-Conditional-Generative-Adversarial-Networks, **StarGAN**: https://github.com/yunjey/stargan, and **AttGAN**: https://github.com/elvisyjlin/AttGAN-PyTorch.



(a) Comparison on the MORPH dataset.

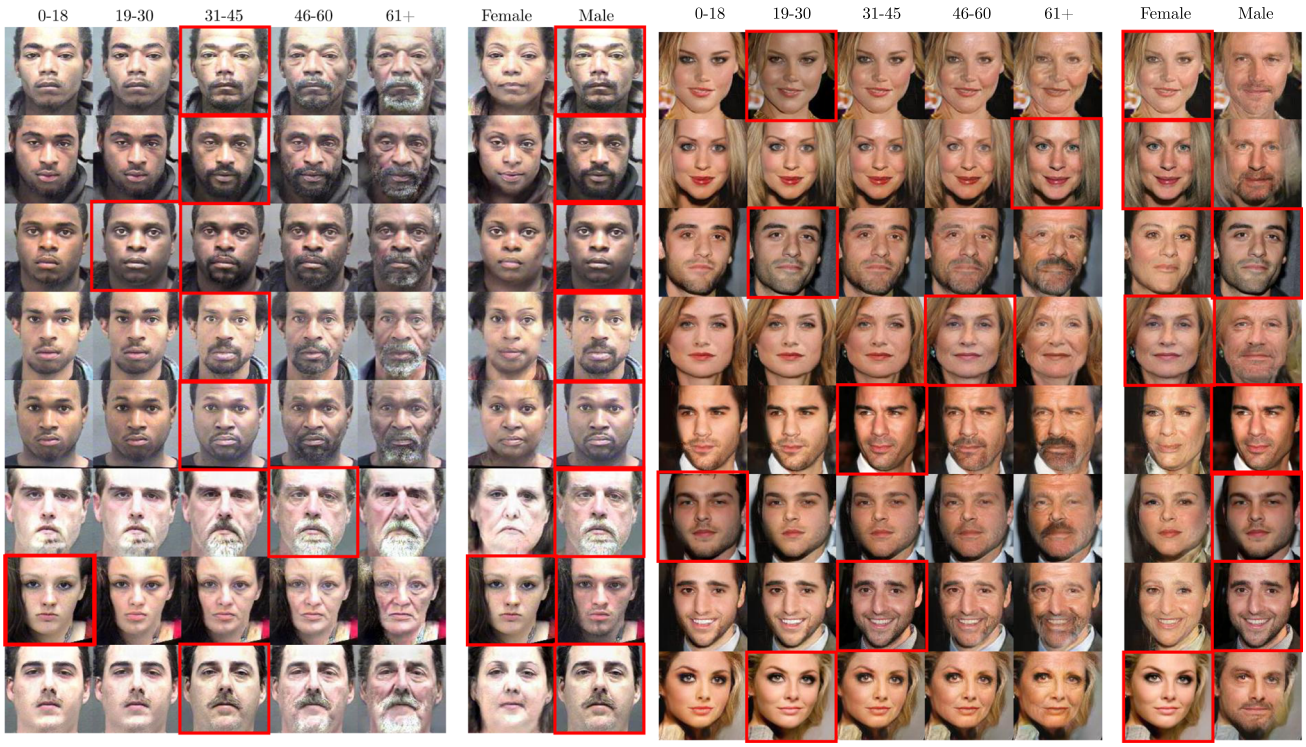


(b) Comparison on the CACD dataset.



(c) Comparison on the UTKFace dataset.

Fig. 2 Multiple attribute transfer on the test sets of MORPH, CACD, and UTKFace. For the two age progression methods (CAAE and IPCGAN), we present the age transfers only. Whilst StarGAN produces images with strong class-discriminative features, many artifacts are present and the sharpness is far inferior to our method's samples. AttGAN fails to produce prominent changes in the input images



(a) Attribute transfer on the MORPH test set

(b) Attribute transfer on the CACD test set

Fig. 3 Multiple demographic attribute transfer on the test sets of MORPH and CACD using our method. Our method is particularly good at aging inputs to the tail ends of the distribution. The red boxes show the input image, in the position of its attribute labels

such as people over 60 years old (see Fig. 2b). Indicatively, only 0.6% of the MORPH and 0.9% of the CACD training sets are over 60 years old. On the other hand, our method is able to synthesize sharp facial images with prominent attribute patterns without sacrificing the photorealism.

Additional results for the task of image translation are presented in Figs. 3a and 3b. The proposed framework is able to transfer distinct patterns for age, gender, and skin tone. In particular, we highlight the transfer of global aging patterns such as wrinkles and hair, as well as features associated with gender, such as jawlines and facial hair. In Fig. 4, we present generated facial images for all attribute combinations on CACD.

The ability of the proposed model to generate images of all demographic subpopulations that exist in the training set–regardless of their representational support–is leveraged in Sect. 4.6 to mitigate the classification bias using data augmentation.

4.4.2 Intra-class Diversity

A useful property of the proposed method lies in its ability to synthesize multiple image variants per class by conditioning on different within-class attribute styles. That is, the network is trained to transfer *specific* attribute styles present in the target image, rather than collapsing to a class-specific pattern. This implies that the synthesized images adapt a given attribute style to the attribute style of the image we condition on, in effect being able to modulate (both attenuate and accentuate) the effect of each target attribute style–thus providing a much more fine-grained control over the synthesis process. For instance in Fig. 5 we demonstrate how an input image can have its attribute content accentuated to different degrees according to the choice of target face. To further demonstrate this point, in Fig. 6 we show that even faces belonging to the oldest age group (over 60 years old) can be translated to look even older by using a target image with more pronounced aging patterns. This property is particularly useful in the case of celebrity datasets (e.g., CACD, KANFace, CELEB-A, LFW) where the appar-

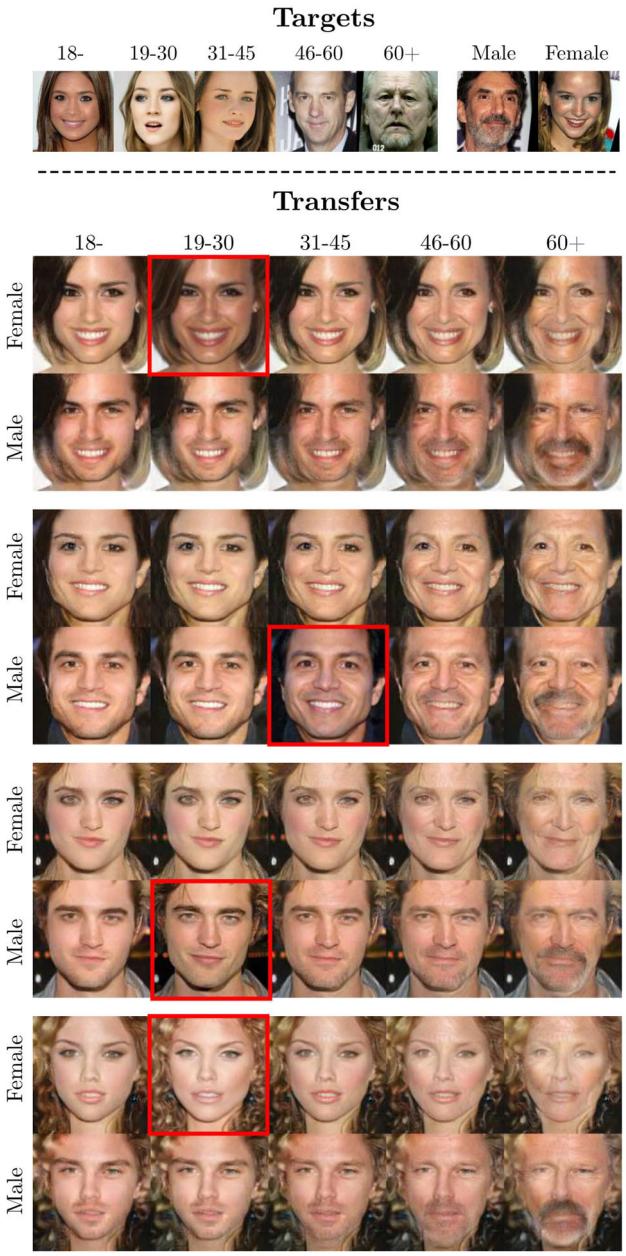


Fig. 4 Translating images from the test set of CACD to all combinations of the dataset's gender and age labels using our method. The red square indicates the input image, and is positioned according to its label

ent age of a face can be significantly lower than its actual age.

4.5 Diversity Enhancement

In this section, we propose to benchmark the diversity-enhancing capabilities of our model with regards to the available demographic attributes. In particular, we translate a test-set of 1000 faces to all attribute classes and calculate the diversity metrics of the augmented set, as proposed in Merler et al. (2019). We measure the Shannon H (ShH) and Simpson D (SiD) diversity indices and the Shannon E (ShE) and Simpson E (SiE) evenness indices. The metrics are calculated as follows:

$$\begin{aligned} &\text{Shannon}: \quad H = -\sum_1^S p_i \ \ln(p_i), \quad E = \frac{H}{\ln(S)} \\ &\text{Simpson}: \quad D = \frac{1}{\sum_1^S p_i^2}, \qquad\qquad E = \frac{D}{S}, \end{aligned}$$

where S denotes the number of classes and p_i is the probability of each class. Higher diversity indices indicate a more diverse dataset, while evenness indices closer to 1 indicate a label distribution that is closer to uniform. The age and gender labels for each image are obtained using the Face++ public API,[3] while the skin tone labels are obtained using the Clarifai[4] API.

In the case of the *age* attribute, we adopt the standard protocol as described in the age progression literature (Yang et al. 2018; Wang et al. 2016; Yang et al. 2016), and translate only faces from the youngest age group (under 18 years old) to the rest of the age groups; that is, we perform face aging.

We compare the diversity metrics of the augmented sets to those of the original test images (GT) and present the results in Tables 1, 2 and 3. Along with our method, both StarGAN and AttGAN are capable of synthesizing a dataset with a distribution of labels close to uniform in the case of the binary attributes skin tone and gender. For the more challenging age attribute however, the superiority of our method and modelling choice is evident, with our augmented datasets' age labels being much more evenly spread than those of the baselines, especially the age progression ones.

This is particularly pronounced in the case of the biased MORPH test-set (Table 1), where the difference between the ground truth images and the synthetic ones is significant. Our model consistently outperforms the baselines for all attributes in both datasets.

[3] https://www.faceplusplus.com/attributes/.

[4] https://clarifai.com/.

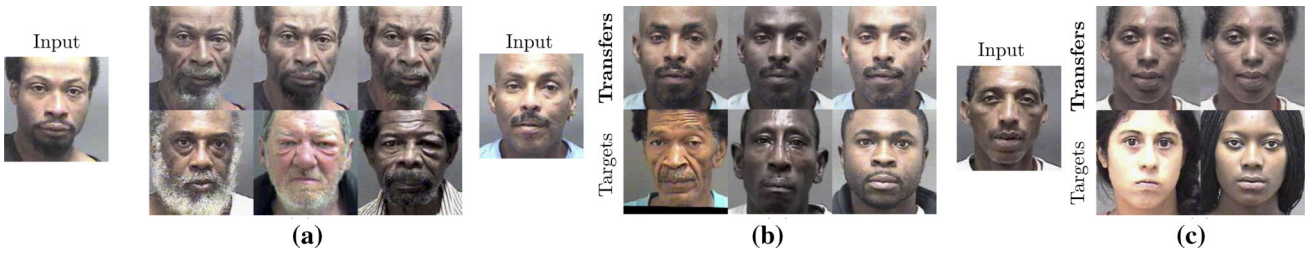


Fig. 5 Our method successfully preserves the intra-class diversity for all of the attributes (for (a) age, (b) skin tone, and (c) gender)

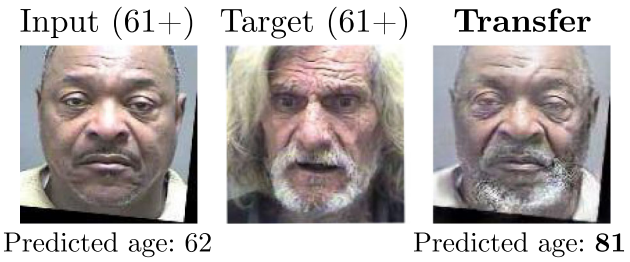


Fig. 6 Intra-class attribute enhancing: the proposed method can accentuate the input class of an attribute label to generate synthetic datasets with distributions with more support on the tails or for the underrepresented classes. The age of the images are evaluated with Face++

4.6 Mitigating Classifier Bias

In this section, we investigate the effect that a lack of diversity entails with respect to model performance. To this end, we fine-tune and test a state-of-the-art gender recognition model (Rothe et al. 2018) on the MORPH and KANFace datasets. By measuring the True Positive Rate (TPR) for each demographic subpopulation, we are able to uncover the bias of the model with regards to age and skin tone. In particular, Figs. 7 and 8 show that the model trained on MOPRH is biased against dark-skinned females (TPR of 40% for dark-skinned females over 60 years old), while the model trained on KANFace is biased against male faces under 18 (TPR of 64%). Age and skin tone bias has been studied for the task of face recognition in Nagpal et al. (2019), as well as in human perception in psychology literature (Schaich et al. 2016; Bothwell et al. 1989).

We propose to mitigate the bias in both models by augmenting the training set using the proposed image translation method. In particular, we train the models on the diverse augmented sets and evaluate the fairness of the trained classifiers. The MORPH dataset is augmented using the models trained on MORPH in Sect. 4.4.1. For the KANFace dataset we use the models trained on CACD.

Of the various fairness metrics discussed in Sect. 2.3, we opt to quantify fairness using Equality of Opportunity (EO) (Hardt et al. 2016), which is defined as the difference in TPRs between the subpopulations.In particular, we report the EO score between each *(age, gender)* class for KANFace and *(age, gender, skin tone)* for MORPH.

We present the TPRs (↑) and EO (↓) on the ground-truth and synthetic test-sets in Figs. 8 and 7 for MORPH and KANFace respectively. Despite StarGAN's images having strong class-discriminative features as established in Sect. 4.4, we demonstrate here that training on their artifact-ridden images leads to even more biased classifiers. Similarly, whilst training on AttGAN-generated training sets can improve the TPR for dark-skinned women over 60 years old by 20% for MORPH, the dataset bias is exacerbated in KANFace when training on augmentations from this model (compared to our augmentations that lead to almost half the EO for males under 18).

Overall, the results indicate that the proposed framework is the only method able to mitigate all of the demonstrated dataset biases in generating more diverse data that can be used to train fairer classifiers.

Table 1 Diversity metrics on the augmented subset of MORPH's test set

	Age ShH	Age ShE	Age SiD	Age SiE	Gender ShH	Gender ShE	Gender SiD	Gender SiE	Skin Tone ShH	Skin Tone ShE	Skin Tone SiD	Skin Tone SiE
GT	–	–	–	–	0.36	0.52	1.26	0.63	0.48	0.70	1.44	0.72
IPCGAN	1.07	0.66	2.27	0.45	–	–	–	–	–	–	–	–
CAAE	0.91	0.57	2.11	0.42	–	–	–	–	–	–	–	–
StarGAN	1.38	0.86	3.67	0.73	0.690	0.996	1.980	0.994	**0.69**	**0.99**	**1.99**	**0.99**
AttGAN	1.29	0.80	3.23	0.65	0.55	0.79	1.57	0.78	**0.69**	**0.99**	**1.99**	**0.99**
Ours	**1.42**	**0.88**	**3.70**	**0.74**	**0.692**	**0.999**	**1.998**	**0.999**	**0.69**	**0.99**	**1.99**	**0.99**

We calculate the indices for the age attribute by age progressing all under 18 s in the test set following the standard practice in the age progression literature (Liu et al. 2019; Wang et al. 2018). For this reason, we dont report the age diversity of the ground-truth images. Our method strictly outperforms or is near-perfect for all metrics and attributes

4.6.1 Comparison to Debiasing Methods

In order to further showcase the efficacy of the proposed framework in bias mitigation, we provide a thorough comparison with 7 state-of-the-art debiasing methods, showcasing the results in Fig. 9. The chosen baselines use different approaches to fairness. Georgopoulos et al. (2020) introduce a method to debias pretrained network embeddings by decomposing them into task specific and protected attribute representations. Similarly, adversarial learning is used to remove the sensitive information from the learned representation in Alvi et al. (2018), Kim et al. (2019), Zhang et al. (2018). An autoencoder that translates the input into fair images is proposed in Quadrianto et al. (2019).[5] Lastly, Wang et al. (2020) use both a domain independent (IND) and a domain discriminative (DISC) approach to fair classification.

Most of these approaches are proposed for binary protected attributes, however our method is more general in affording the ability to handle the case of multiple protected attributes (e.g., age group). The results in Fig. 9a highlight the ability of the proposed method to mitigate the age bias of the gender classifier, especially in the case of male faces under 18 years old (age group a0) in the KANFace dataset. In particular, our framework achieves significantly lower EO (0.19–0.22 EO for age group 0) with Georgopoulos et al. (2020) being the second best (0.2–0.26 EO for age group 0).

We conduct a similar experiment on MOPRH, where the sensitive attribute is binary (i.e., skin tone). The results in Fig. 9b show that our method can produce competitive results. However, our method doesn't achieve the lowest EO, which can be attributed to the fact that the rest of the baselines were designed to tackle binary protected attributes. In particular (Wang et al. 2020) IND and (Quadrianto et al. 2019) achieve the most fair classification results, with our method being a close second.

[5] For this method, we find that penalizing the reconstruction in the image space performs significantly better than doing so in the feature space (as proposed in the original paper). Therefore, we implement the model as such.

Table 2 Diversity metrics on the augmented subset of CACD's test set

	Age ShH	Age ShE	Age SiD	Age SiE	gender ShH	gender ShE	gender SiD	gender SiE
GT	–	–	–	–	0.6928	0.9995	1.9986	0.9993
IPCGAN	1.16	0.72	2.87	0.57	–	–	–	–
CAAE	0.90	0.56	2.16	0.43	–	–	–	–
StarGAN	**1.39**	**0.86**	3.80	0.76	0.6906	0.9963	1.9899	0.9950
AttGAN	1.12	0.69	2.68	0.54	0.6922	0.9987	1.9963	0.9982
Ours	**1.39**	**0.86**	**3.89**	**0.78**	**0.6931**	**0.9999**	**1.9997**	**0.9999**

We calculate the indices for the age attribute by age progressing all under 18 s in the test set following the standard practice in the age progression literature (Liu et al. 2019; Wang et al. 2018). For this reason, we dont report the age diversity of the ground-truth images

4.7 Limitations of the Framework

Due to the generative nature of the proposed framework, our method suffers from its dependence to external data. We investigate this inherent shortcoming of our method by conducting two experiments, which are presented as follows.

Training on limited data Firstly, we investigate the impact of the size of the training set of the GAN on the bias of the trained classifier. Concretely, we show in Fig. 10 the results of training a classifier with our synthetic data, when our GAN is trained on training subsets of different sizes. For each model, we report the equality of opportunity for the most underrepresented class, i.e., male faces under 18 on KANFace. We notice that when our model is trained on less than 25% of the training images, the bias of the trained classifier is even magnified. It should be noted that this problem is not specific to our method, but one that will plague all generative model-based bias mitigation methods via augmentation.

Augmentation with real data Since the use of the proposed framework assumes the existence of an external training set, we explore the option of augmenting the training set of the biased classifier using real images from the external set. In particular, we use the CACD dataset as the external set, and try to flatten the distribution of the training set of the classifier (i.e., KANFace). However, the support for each class in the external set is not sufficient to do so in the extreme cases of people under 18 and over 60 years old. The resulting training distributions along with equality of opportunity scores are presented in Fig. 11. The results highlight the advantage of using the proposed data augmentation method, instead of directly training on the external data.

Learning the real distribution of face images is a task of high data and computational complexity (Arora and Zhang 2017). Therefore, generative models have to rely on their inductive bias in order to generalize on unobserved modes of variation. However, any finite dataset is inherently biased, and training a generative model on a biased dataset can even exacerbate this bias. Indicatively, GANs and VAEs are not able to generate images of unobserved attribute combinations (Zhao et al. 2018). Different approaches for debiased generation have been proposed in the literature and include importance reweighting (Grover et al. 2019a, b) and modeling the multiplicative interactions (Georgopoulos et al. 2020). Therefore, the proposed and any generative model-based approach to data augmentation should be utilized with caution regarding the distribution of the training set.

5 Conclusions

In this paper, we proposed a style-based neural data augmentation framework, that can be used to enhance the demographic diversity of a given dataset. To this end, we introduced a novel style transfer method, that is able to simultaneously transfer multiple facial demographic attributes. Contrary to recent work in multi-attribute face translation, the proposed framework leverages attribute-specific demographic style to facilitate image translation, rather than class-labels. In order to mix the style information of multiple attributes, we further introduce a multilinear extension to AdaIN, that fuses the different styles by modeling their multiplicative interactions.

Table 3 Diversity metrics on the augmented subset of UTKFace's test set

	Age ShH	Age ShE	Age SiD	Age SiE	Gender ShH	Gender ShE	Gender SiD	Gender SiE	Skin Tone ShH	Skin Tone ShE	Skin Tone SiD	Skin Tone SiE
GT	–	–	–	–	**0.69**	**0.99**	**1.99**	**0.99**	0.57	0.82	1.61	0.81
IPCGAN	1.40	0.87	3.69	0.74	–	–	–	–	–	–	–	–
CAAE	1.48	0.92	4.06	0.81	–	–	–	–	–	–	–	–
StarGAN	1.46	0.91	3.82	0.76	**0.69**	**0.99**	**1.99**	0.98	0.68	0.99	1.96	0.98
AttGAN	1.32	0.82	3.30	0.66	**0.69**	**0.99**	**1.99**	**0.99**	0.58	0.83	1.63	0.82
Ours	**1.53**	**0.95**	**4.35**	**0.87**	**0.69**	**0.99**	**1.99**	**0.99**	**0.69**	**0.999**	**1.997**	**0.999**

We calculate the indices for the age attribute by age progressing all under 18 s in the test set following the standard practice in the age progression literature (Liu et al. 2019; Wang et al. 2018). For this reason, we dont report the age diversity of the ground-truth images

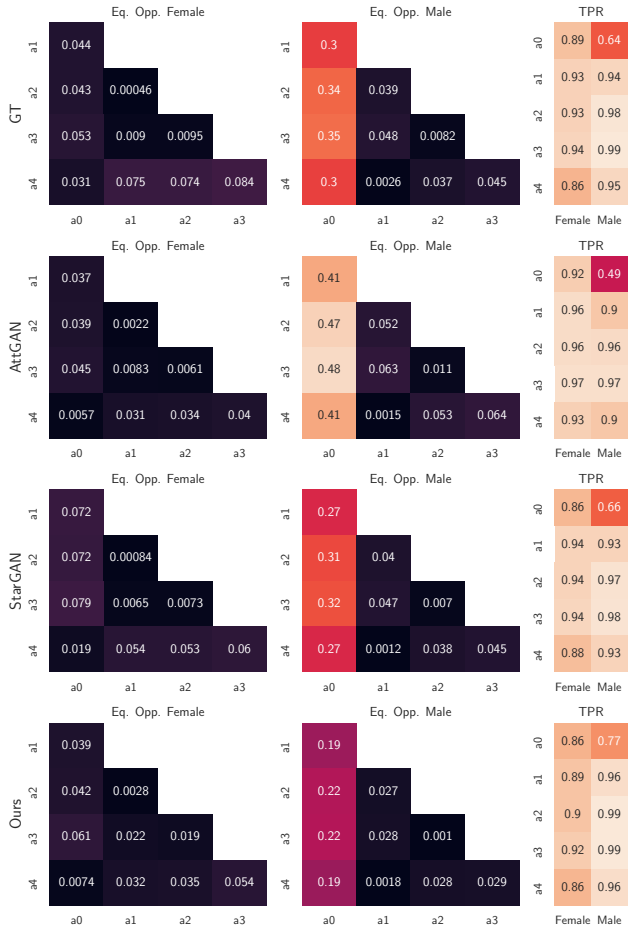


Fig. 7 Equal Opportunity (↓) and TPRs (↑) for the two subpopulations for gender recognition on KANFace's test set. Trained on the augmentations from the proposed method, the classifiers are notably less biased against young males. 'ai' denotes age class i

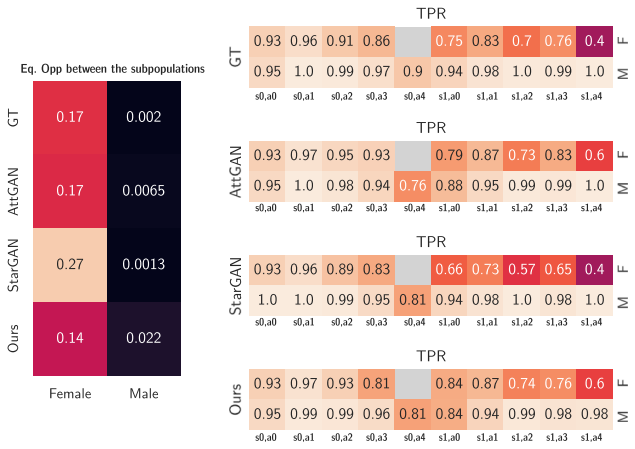
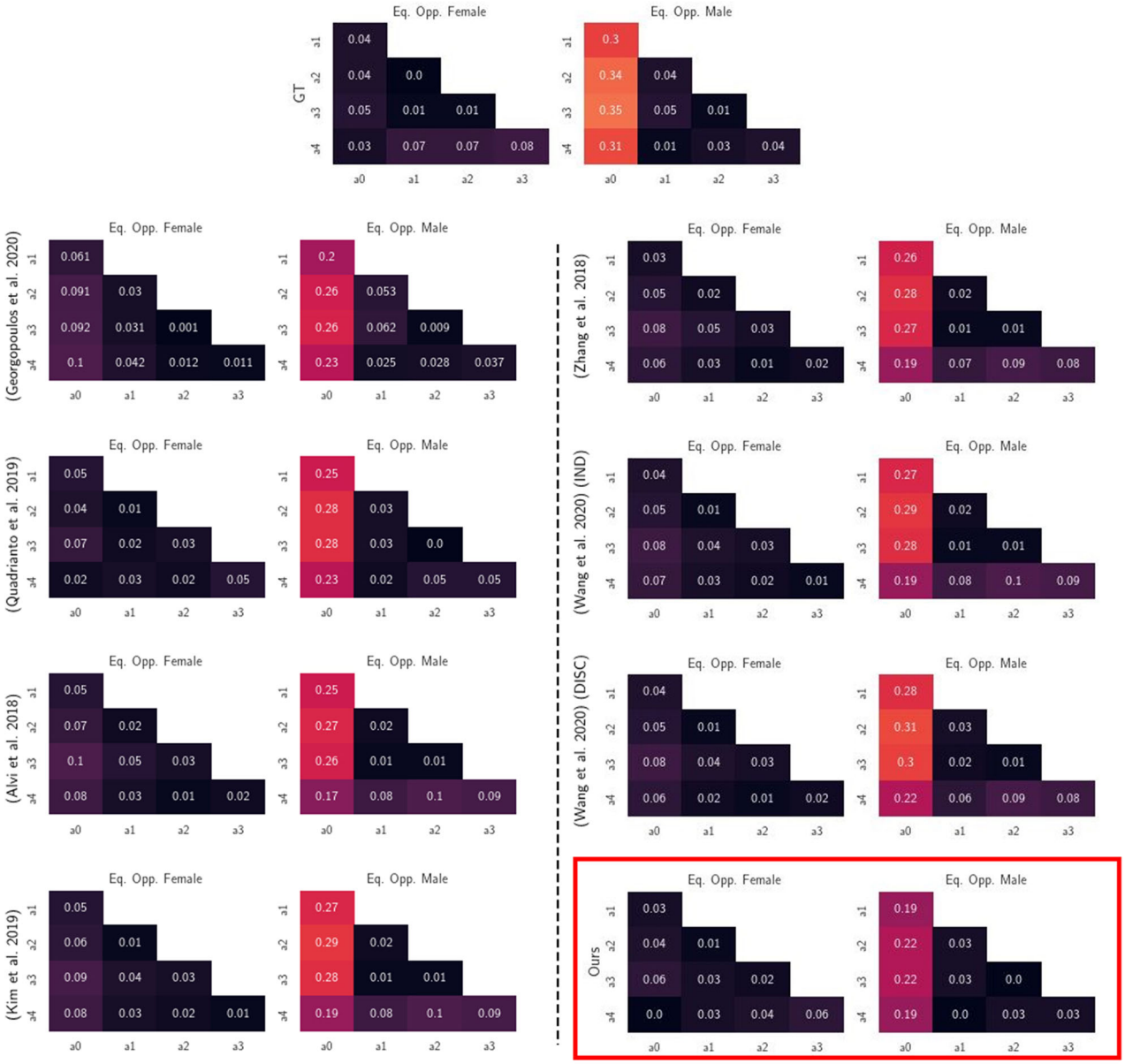
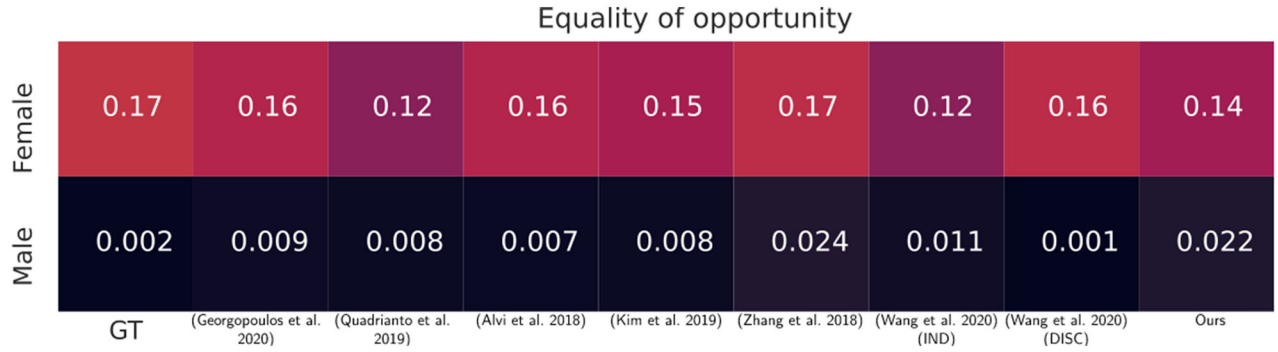


Fig. 8 Equal Opportunity (↓) and TPRs (↑) for the two subpopulations for gender recognition on MORPH's test set. Trained on our augmentations, the classifiers are notably less biased against dark-skinned females of all ages. We note that there are no light-skinned females aged over 60 in the test set (s0, a4) and hence their entries have been left blank. 'ai' denotes age class i, while s0 and s1 refer to light- and dark-skin tones respectively. Full results for all combinations of the labels can be found in the supplementary material



(a) EO (for the age attribute) for the two subpopulations for gender recognition on KANFace's test set. The red rectangle highlights our method's results.



(b) EO (for the skin tone attribute) for the two subpopulations for gender recognition on MORPH's test set.

Fig. 9 Equal Opportunity (EO) (↓) for the subpopulations for gender recognition on MORPH and KANFace's test set. Our method produces classifiers that compete with all 7 state-of-the-art bias mitigation methods

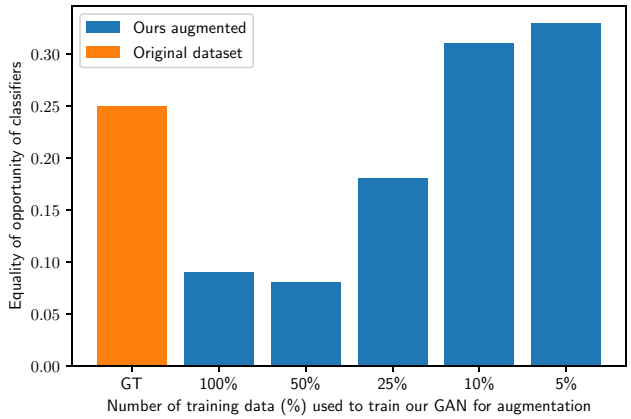


Fig. 10 Equality of Opportunity of classifiers trained on the ground-truth biased training data and the augmented KAN-Face training sets

Our style transfer framework is evaluated against baseline attribute transfer models (StarGAN and AttGAN), as well as GAN-based age progression methods (CAAE and IPCGAN) in a series of qualitative and quantitative experiments. In particular, we demonstrate that the proposed model is able to realistically transfer age, gender, and skin tone on the CACD and MORPH datasets, even for undrerepresented classes (such as images of people over 60 or under 18 years old). We further evaluate the diversity-enhancing capabilities of the models by measuring the diversity [as proposed in Merler et al. (2019)] of the augmented test sets. By augmenting the test set using our method, we are able to achieve the most evenly spread distribution of age, gender, and skin tone predictions.

Lastly, we quantify the effect of biased datasets in the fairness of face analysis models. Focusing on gender recognition, we showcase how our framework can be used to mitigate existing age and skin tone bias in a state-of-the-art model (Rothe et al. 2018) on the MORPH and KANFace datasets. By measuring equality of opportunity, we show that the model is more fair when trained on the augmentations produced by our model. At the same time, the results in Figs. 7 and 8 indicate that training on non-diverse or non-photorealistic synthetic images can even deteriorate the fairness of a pretrained classifier.

In our future work, we aim to extend our style-transfer framework to be able to handle heavily biased datasets (e.g., missing classes). Furthermore, we plan to investigate different models that are not so dependent on large annotated training sets, e.g., non adversarial frameworks. Another direction is to extend the method to handle an arbitrary number of attributes, without the restrictions imposed by the size of the tensors. This can be achieved by using coupled tensor decompositions that utilize parameter sharing. Lastly, we plan to extend our method to better account for attributes for which a discrete categorization is not so applicable. In particular, we plan to address bias beyond the traditional binary

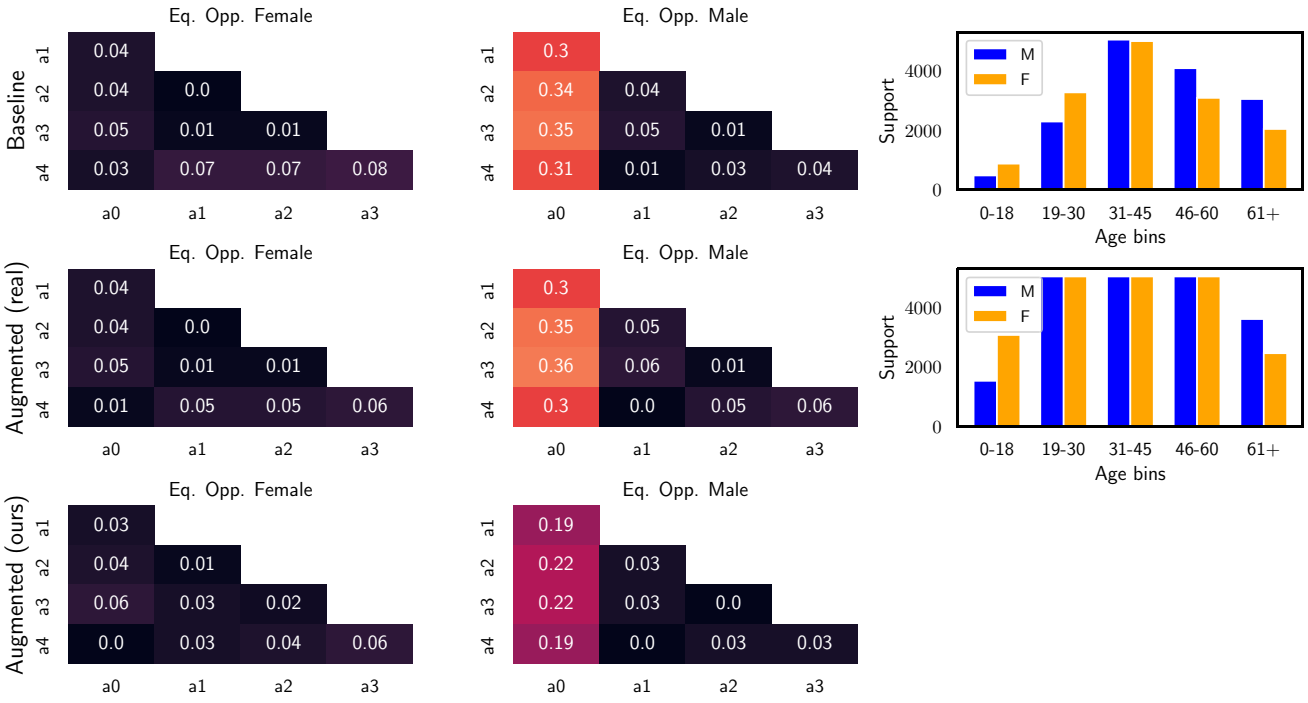


Fig. 11 Equality of Opportunity scores (two left-most columns) of the classifiers trained using the ground-truth data (first row), the data augmented with real images from a different dataset (middle row), and with our synthetic images (bottom row). The right-most column shows the support of the datasets used to train the classifiers for each row

gender paradigm to better reflect the full spectrum of gender identity.

References

Alvi, M., Zisserman, A., & Nellåker, C. (2018). Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European conference on computer vision (ECCV)* (p. 0)

Arjovsky, M., Chintala, S., & Bottou, L. (2017). *Wasserstein gan*

Arora, S., Zhang, Y. (2017). *Do GANs actually learn the distribution? An empirical study*. arXiv preprint arXiv:1706.08224

Bothwell, R. K., Brigham, J. C., & Malpass, R. S. (1989). Cross-racial identification. *Personality and Social Psychology Bulletin*, *15*(1), 19–25.

Brock, A., Donahue, J., & Simonyan, K. (2018). *Large scale GAN training for high fidelity natural image synthesis*.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77–91).

Carroll, J. D., & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of Eckart–YoungÂ decomposition. *Psychometrika*, *35*(3), 283–319.

Chen, B.C., Chen, C.S., & Hsu, W.H. (2014). Cross-age reference coding for age-invariant face recognition and retrieval. In *Proceedings of the European conference on computer vision (ECCV)*.

Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *2018 IEEE/CVF conference on computer vision and pattern recognition*. https://doi.org/10.1109/cvpr.2018.00916.

Dantcheva, A., Elia, P., & Ross, A. (2015). What else does your biometric data reveal? A survey on soft biometrics. *IEEE Transactions on Information Forensics and Security*, *11*(3), 441–467.

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. http://archive.ics.uci.edu/ml.

Duong, C.N., Luu, K., Quach, K.G., Nguyen, N., Patterson, E., Bui, T.D., & Le, N. (2019). Automatic face aging in videos via deep reinforcement learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 10013–10022).

Edwards, H., & Storkey, A. (2015). *Censoring representations with an adversary*. arXiv preprint arXiv:1511.05897.

Fu, Y., Guo, G., & Huang, T. S. (2010). Age synthesis and estimation via faces: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(11), 1955–1976.

Gatys, L.A., Ecker, A.S., & Bethge, M. (2015). *A neural algorithm of artistic style*. arXiv:1508.06576.

Georgopoulos, M., Chrysos, G., Pantic, M., & Panagakis, Y. (2020). Multilinear latent conditioning for generating unseen attribute combinations. In *International conference on machine learning*

Georgopoulos, M., Oldfield, J., Nicolaou, M.A., Panagakis, Y., & Pantic, M. (2020). Enhancing facial data diversity with style-based face aging. In *2020 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), IEEE* (pp. 66–74), Seattle, WA, USA. https://doi.org/10.1109/CVPRW50498.2020.00015. https://ieeexplore.ieee.org/document/9150573/.

Georgopoulos, M., Panagakis, Y., & Pantic, M. (2018). Modeling of facial aging and kinship: A survey. *Image and Vision Computing*, *80*, 58–79.

Georgopoulos, M., Panagakis, Y., & Pantic, M. (2020). Investigating bias in deep face analysis: The kanface dataset and empirical study. *Image and Vision Computing*. https://doi.org/10.1016/j.imavis.2020.103954.

Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative adversarial networks*

Grover, A., Choi, K., Shu, R., & Ermon, S. (2019a). *Fair generative modeling via weak supervision*. arXiv preprint arXiv:1910.12008.

Grover, A., Song, J., Kapoor, A., Tran, K., Agarwal, A., Horvitz, E.J., & Ermon, S. (2019b). Bias correction of learned generative models using likelihood-free importance weighting. In *Advances in neural information processing systems* (pp. 11058–11070).

Hardt, M., Price, E., Srebro, N., & et al. (2016). Equality of opportunity in supervised learning. In *Advances in neural information processing systems* (pp. 3315–3323).

Harshman, R. A., et al. (1970). *Foundations of the parafac procedure: Models and conditions for an " explanatory" multimodal factor analysis*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.90.

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2017). *Arbitrary facial attribute editing: Only change what you want*. arXiv:1711.10678.

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2019). Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, *28*(11), 54645478.

Hendricks, L.A., Burns, K., Saenko, K., Darrell, T., & Rohrbach, A. (2018). Women also snowboard: Overcoming bias in captioning models. In *European conference on computer vision* (pp. 793–811). Springer.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In: *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–16).

Huang, G.B., Mattar, M., Berg, T., & Learned-Miller, E. (2008). *Labeled faces in the wild: A database forstudying face recognition in unconstrained environments*.

Huang, X., & Belongie, S. (2017). *Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization*. arXiv:1703.06868.

Huang, X., Liu, M.Y., Belongie, S., & Kautz, J. (2018). *Multimodal unsupervised image-to-image translation*. arXiv:1804.04732.

Inoue, H. (2018). *Data augmentation by pairing samples for images classification*. arXiv preprint arXiv:1801.02929

Isola, P., Zhu, J.Y., Zhou, T., & Efros, A.A. (2017). Image-to-image translation with conditional adversarial networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/cvpr.2017.632.

Jackson, P. T., Abarghouei, A. A., Bonner, S., Breckon, T. P., & Obara, B. (2019). Style augmentation: data augmentation via style randomization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Jayakumar, S.M., Menick, J., Czarnecki, W.M., Schwarz, J., Rae, J., Osindero, S., Teh, Y.W., Harley, T., & Pascanu, R. (2020). Multiplicative interactions and where to find them. In *International conference on learning representations*.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2017). *Progressive growing of GANs for improved quality, stability, and variation*.

Karras, T., Laine, S., & Aila, T. (2019). *A style-based generator architecture for generative adversarial networks*. arXiv:1812.04948.

Kim, B., Kim, H., Kim, K., Kim, S., & Kim, J. (2019). Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 9012–9020).

Kim, J., Kim, M., Kang, H., & Lee, K. (2020). *U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation*. arXiv:1907.10830.

Kingma, D., & Ba, J. (2014). *Adam: A method for stochastic optimization. International Conference on Learning Representations*.

Kolda, T. G. (2006). *Multilinear operators for higher-order decompositions*. Technical Reports, Sandia National Laboratories.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, *51*(3), 455–500.

Kuhlman, C., Jackson, L., & Chunara, R. (2020). *No computation without representation: Avoiding data and algorithm biases through diversity*. arXiv preprint arXiv:2002.11836.

Lanitis, A. (2002). *FG-NET Aging Database*.

Li, M., Zuo, W., & Zhang, D. (2016). *Deep identity-aware transfer of facial attributes*.

Li, S., & Deng, W. (2020). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*.

Lim, J.H., & Ye, J.C. (2017). *Geometric GAN*.

Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., & Kautz, J. (2019). *Few-shot unsupervised image-to-image translation*. arXiv: 1905.01723v2.

Liu, Y., Li, Q., Sun, Z., & Tan, T. (2019). *A3gan: An attribute-aware attentive generative adversarial network for face aging*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of international conference on computer vision (ICCV)*.

Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., & Gool, L.V. (2018). *Exemplar guided unsupervised image-to-image translation with semantic consistency*.

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018). *Learning adversarially fair and transferable representations*. arXiv preprint arXiv:1802.06309.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., & Smolley, S.P. (2017). Least squares generative adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)*. https://doi.org/10.1109/iccv.2017.304.

Masi, I., Wu, Y., Hassner, T., & Natarajan, P. (2018) Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI), IEEE* (pp. 471–478).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). *A survey on bias and fairness in machine learning*. arXiv preprint arXiv:1908.09635.

Merler, M., Ratha, N., Feris, R. S., & Smith, J. R. (2019). *Diversity in faces*.

Mescheder, L., Geiger, A., & Nowozin, S. (2018). *Which Training Methods for GANs do actually Converge?* arXiv:1801.04406 [cs].

Nagpal, S., Singh, M., Singh, R., Vatsa, M., & Ratha, N. (2019). *Deep learning for face recognition: Pride or prejudiced?* arXiv preprint arXiv:1904.01219.

Ng, C.B., Tay, Y.H., & Goi, B.M. (2012). *Vision-based human gender recognition: A survey*. arXiv preprint arXiv:1204.1611.

Odena, A., Olah, C., & Shlens, J. (2017). Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 2642–2651). JMLR. org.

Park, T., Liu, M.Y., Wang, T.C., & Zhu, J.Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In: *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*. https://doi.org/10.1109/cvpr.2019.00244.

Perarnau, G., van de Weijer, J., Raducanu, B., & Ivarez, J.M. (2016). *Invertible conditional GANs for image editing*.

Perez, L., & Wang, J. (2017). *The effectiveness of data augmentation in image classification using deep learning*. arXiv preprint arXiv:1712.04621.

Quadrianto, N., Sharmanska, V., & Thomas, O. (2018). *Discovering fair representations in the data domain*.

Quadrianto, N., Sharmanska, V., & Thomas, O. (2019). Discovering fair representations in the data domain. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8227–8236).

Radford, A., Metz, L., & Chintala, S. (2015). *Unsupervised representation learning with deep convolutional generative adversarial networks*.

Raji, I.D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *Proceedings of the 2019 AAAI/ACM conference on AI, ethics, and society* (pp. 429–435).

Ramanathan, N., Chellappa, R., Biswas, S., et al. (2009). *Age progression in human faces: A survey*. *Visual Languages and Computing, 15*, 3349–3361.

Ricanek, K., & Tesafaye, T. (2006). Morph: A longitudinal image database of normal adult age-progression. In *7th international conference on automatic face and gesture recognition (FGR06)* (pp. 341–345). https://doi.org/10.1109/FGR.2006.78.

Rothe, R., Timofte, R., & Gool, L. V. (2018). Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, *126*(2–4), 144–157.

Salimans, T., Goodfellow, I. J., Zaremba, W., Cheung, V., Radford, A., & Chen, X. (2016). Improved techniques for training GANs. In *CoRR*. arXiv:1606.03498.

Sandfort, V., Yan, K., Pickhardt, P. J., & Summers, R. M. (2019). Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in CT segmentation tasks. *Scientific Reports*, *9*(1), 1–9.

Sattigeri, P., Hoffman, S.C., Chenthamarakshan, V., & Varshney, K.R. (2018). *Fairness GAN*.

Schaich, A., Obermeyer, S., Kolling, T., & Knopf, M. (2016). An own-age bias in recognizing faces with horizontal information. *Frontiers in Aging Neuroscience*, *8*, 264.

Serna, I., Morales, A., Fierrez, J., Cebrian, M., Obradovich, N., & Rahwan, I. (2019). *Algorithmic discrimination: Formulation and exploration in deep learning-based face biometrics*. arXiv preprint arXiv:1912.01842.

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60.

Tang, H., Liu, H., Xu, D., Torr, P.H.S., & Sebe, N. (2019). *Attention-gan: Unpaired image-to-image translation using attention-guided generative adversarial networks*.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare), IEEE* (pp. 1–7).

Wang, M., Deng, W., Hu, J., Tao, X., & Huang, Y. (2019). Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE international conference on computer vision*, pp. 692–702.

Wang, W., Cui, Z., Yan, Y., Feng, J., Yan, S., Shu, X., & Sebe, N. (2016). Recurrent face aging. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2378–2386).

Wang, Z., Qinami, K., Karakozis, I.C., Genova, K., Nair, P., Hata, K., & Russakovsky, O. (2020). Towards fairness in visual recognition: Effective strategies for bias mitigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8919–8928).

Wang, Z. X., Tang, W. L., & Gao, S. (2018). Face aging with identity-preserved conditional generative adversarial networks. In *2018 IEEE conference on computer vision and pattern recognition (CVPR)*.

Yang, H., Huang, D., Wang, Y., & Jain, A. K. (2018). Learning face age progression: A pyramid architecture of GANs. In *2018 IEEE/CVF conference on computer vision and pattern recognition*. https://doi.org/10.1109/cvpr.2018.00011.

Yang, H., Huang, D., Wang, Y., Wang, H., & Tang, Y. (2016). Face aging effect simulation using hidden factor analysis joint sparse representation. *IEEE Transactions on Image Processing*, *25*(6), 2493–2507.

Yucer, S., Akçay, S., Al-Moubayed, N., & Breckon, T. P. (2020). *Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation*. arXiv preprint arXiv:2004.08945.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 335–340).

Zhang, H., Cisse, M., Dauphin, Y.N., & Lopez-Paz, D. (2017). *mixup: Beyond empirical risk minimization*. arXiv preprint arXiv:1710.09412.

Zhang, Z., Song, Y., & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., & Chang, K. W. (2017). *Men also like shopping: Reducing gender bias amplification using corpus-level constraints*. arXiv preprint arXiv:1707.09457.

Zhao, S., Ren, H., Yuan, A., Song, J., Goodman, N., & Ermon, S. (2018). Bias and generalization in deep generative models: An empirical study. In *Advances in Neural Information Processing Systems* (pp. 10792–10801).

Zheng, X., Chalasani, T., Ghosal, K., Lutz, S., & Smolic, A. (2019). *Stada: Style transfer as data augmentation*. arXiv preprint arXiv:1909.01056.

Zhu, J.Y., Park, T., Isola, P., & Efros, A.A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)*. https://doi.org/10.1109/iccv.2017.244.