Fair Generative Models via Transfer Learning

Christopher T.H.Teo[*], **Milad Abdollahzadeh**[*], **Ngai-Man Cheung**[†]

Singapore University of Technology and Design (SUTD)
christopher_teo@mymail.sutd.edu.sg, {milad_abdollahzadeh, ngaiman_cheung}@sutd.edu.sg

Abstract

This work addresses *fair* generative models. Dataset biases have been a major cause of unfairness in deep generative models. Previous work had proposed to augment *large, biased* datasets with *small, unbiased* reference datasets. Under this setup, a weakly-supervised approach has been proposed, which achieves state-of-the-art quality and fairness in generated samples. **In our work**, based on this setup, we propose a simple yet effective approach. Specifically, first, we propose **fairTL**, a *transfer learning* approach to learn fair generative models. Under fairTL, we pre-train the generative model with the available large, biased datasets and subsequently adapt the model using the small, unbiased reference dataset. Our fairTL can learn expressive sample generation during pre-training, thanks to the large (biased) dataset. This knowledge is then transferred to the target model during adaptation, which also learns to capture the underlying fair distribution of the small reference dataset. Second, we propose **fairTL++**, where we introduce two additional innovations to improve upon fairTL: (i) multiple feedback and (ii) Linear-Probing followed by Fine-Tuning (LP-FT). Taking one step further, we consider an alternative, challenging setup when only a pre-trained (potentially biased) model is available but the dataset used to pre-train the model is *inaccessible*. We demonstrate that our proposed fairTL and fairTL++ remains very effective under this setup. We note that previous work requires access to large, biased datasets and cannot handle this more challenging setup. Extensive experiments show that fairTL and fairTL++ achieve state-of-the-art in both quality and fairness of generated samples. *The code and additional resources can be found at* bearwithchris.github.io/fairTL/

Introduction

Deep generative models such as Generative Adversarial Network (GAN) are an active research area (Goodfellow et al. 2014; Brock, Donahue, and Simonyan 2019; Karras et al. 2020; Ojha et al. 2021). Various GAN-based approaches have achieved outstanding results in many tasks, for example: image synthesis (Karras, Laine, and Aila 2019a; Yu et al. 2019) , image transformation (Wang et al. 2018a) , super-resolution (Lucas et al. 2019; Nasrollahi et al. 2020)

---

[*]Equal Contribution   [†]Corresponding Author

, text-to-image synthesis (Zhang et al. 2017) and anomaly detection (Schlegl et al. 2017; Lim et al. 2018).

In recent times, **fairness in generative models** has attracted increasing attention (Frankel and Vendrow 2020; Choi et al. 2020; Humayun, Balestriero, and Baraniuk 2022; Tan, Shen, and Zhou 2020). It is defined as the *equal representation* (Hutchinson and Mitchell 2019) of some selected *sensitive attribute* (SA). For example, a generative model that has an equal probability of producing male and female samples is fair *w.r.t.* Gender. Generative models have been increasingly adopted in various applications including high-stakes areas such as criminal justice (Jalan et al. 2020) and healthcare (Frid-Adar et al. 2018). This brings about concerns regarding potential biases and unfairness of these models. For example, generative models have been applied in suspect facial profiling (Jalan et al. 2020). In this application, a generative model could result in wrongful incrimination of an individual if the model has biases *w.r.t.* certain SA such as Gender or Race. Furthermore, some generative models have been applied to create data for training downstream models *e.g.* classifiers for disease diagnosis (Frid-Adar et al. 2018). Such biases in generative models can propagate to downstream models, exacerbating the situation.

Dataset biases are a major cause of unfairness in deep generative models. Typically, generative models like GANs are trained in an unsupervised manner to capture the underlying distribution of the given dataset, and then generate new data from the same distribution. It is usually expected that the training dataset is large and unbiased *w.r.t.* SAs. This assumption usually holds true when we follow good practices for data collection, such as protocols adopted by biotech companies, or governmental and international institutions such as the World Bank (Choi et al. 2020; Katal, Wazid, and Goudar 2013; Chizzola, Micheli, and Vingelli 2017). However, these protocols are usually unscalable, and collected *fair datasets are usually small* in size (Choi et al. 2020). Therefore, in order to collect the required large dataset, we usually use alternative sources with a related distribution, such as scrapping images from the internet (Muehlethaler and Albert 2021). Collected data from these *alternative resources are usually biased w.r.t.* SAs (Le Quy et al. 2022; Hwang 2020) , and these biases are easily picked up by the generative models.
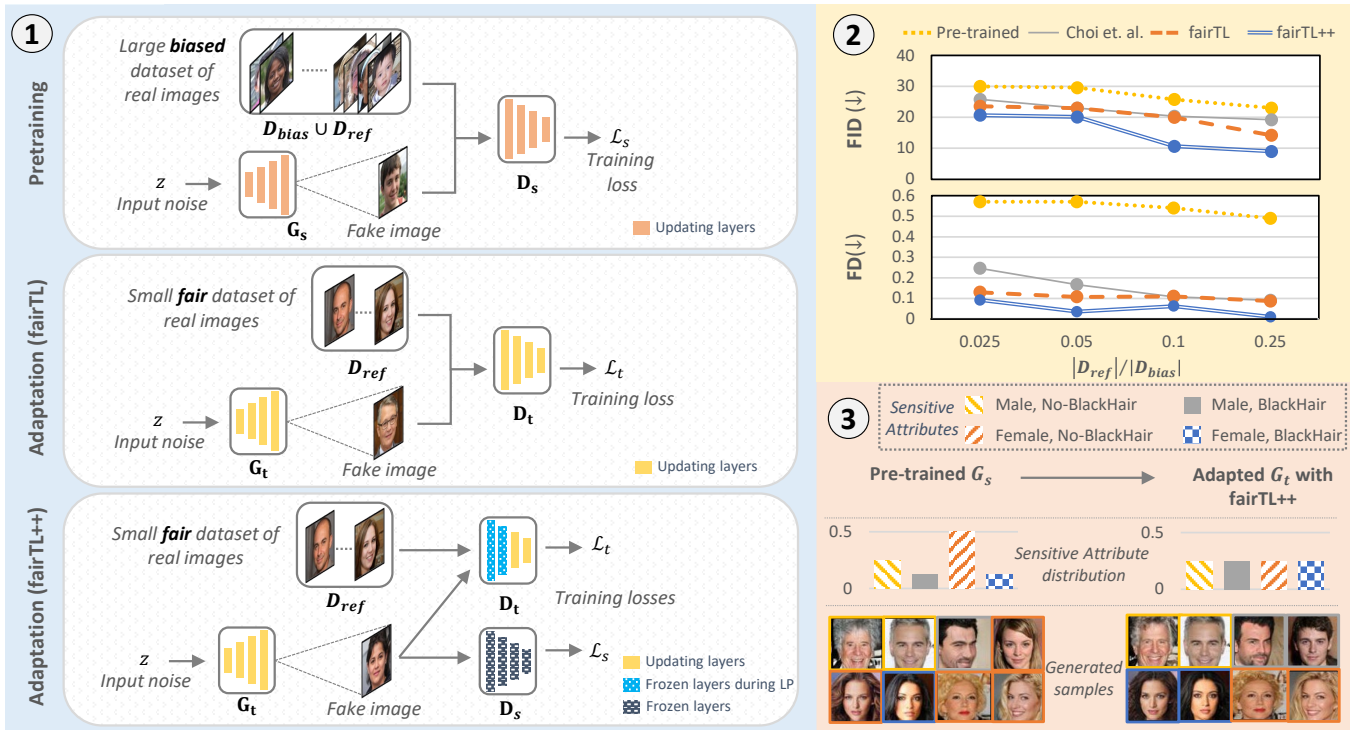
Figure 1: Overview of our work on training fair generative models. ① We train a high-quality generator with fair sensitive attribute (SA) distribution in a two-stage process that we call *fairTL*. In *pre-training*, the GAN learns to generate diverse and high-quality samples from a large but biased dataset. Then, in *adaptation*, the same GAN learns the fair underlying SA distribution from a small reference distribution, D_{ref}. To improve adaptation, we introduced a second variation called *fairTL++*, which includes an additional source of feedback (D_s) and a Linear-Probing step prior to Fine-Tuning. Pre-training step is same for both fairTL and fairTL++. ② Our results from training a GAN with a large biased dataset, D_{bias}, with SA distribution of 90% females and 10% males and a small fair dataset, D_{ref}, with varying $|D_{ref}|$ denoted by $perc = |D_{ref}|/|D_{bias}|$, from celebA (Liu et al. 2015). We compare four approaches *1) pre-training*, 2) (Choi et al. 2020): the SOTA technique, 3) our proposed *fairTL* and 4) *fairTL++*. We then measure the Fréchet Inception Distance (FID) and Fairness Discrepancy (FD) of the four models. A smaller FID indicates better quality and smaller FD indicates better fairness. Without consideration of fairness, the *pre-trained* setup expresses a large bias. Choi *et al.* then significantly improves on this but brings about diminishing quality and fairness as $|D_{ref}|$ becomes smaller. Meanwhile, our proposed method demonstrates greater robustness under these same limitations, achieving SOTA results in both FID and FD. ③ We illustrate the improved fairness on multiple SA {Gender,Blackhair} during the adaptation stage. To do this, we utilize a fixed noise vector, z to sample from both the pre-trained and fairTL++ models. Observe how the majority-represented SA are adapted to the minority-represented SA, thereby improving the SA distribution.

To prevent the biased dataset from harming the fairness of the generative model, the large biased dataset D_{bias} can be augmented with a small fair (*w.r.t.* some specific SAs) dataset D_{ref}, as proposed in (Choi et al. 2020). In this setup, the main idea is that the generative model can learn expressive representation using D_{bias}, while mitigating the bias with D_{ref}. Note that in their setup, neither datasets are labeled *w.r.t.* SAs, and the size of the fair dataset can be much smaller than the biased dataset. For example, $|D_{ref}|$ could be 2.5% of $|D_{bias}|$.

In our work, initially, we follow the setup as in (Choi et al. 2020), and propose a simple *transfer learning* approach for bias mitigation. Specifically, we first propose fair transfer learning (**fairTL**) where we pre-train the generative model using the large biased dataset to learn expressive sample generation. Subsequently, on top of the learned

expressive knowledge, we adapt the model using the small fair dataset to capture the fair SA adaptation. We show that this simple transfer learning approach can be considered as a *strong baseline* for training a fair generative model via transfer learning (Figure 1). However, as D_{ref} is small, the fine-tuning on D_{ref} is susceptible to mode collapse (Mo, Cho, and Shin 2020; Li et al. 2020). Hence, as we adapt the model to learn a fairer SA distribution, it is important to preserve the general knowledge efficiently. To this aim, we propose **fairTL++** where we include two additional improvements upon fairTL: i) *multiple feedback approach*, and ii) *Linear-Probing before Fine-Tuning* (LP-FT). We find that these two innovations can achieve noticeable gain when applied to fairTL individually. Furthermore, when applied together, we are able to achieve significant gain in sample quality and fairness over previous work (Choi et al. 2020). In partic-

ular, fairTL and fairTL++ differentiates itself by *removing the need for a density ratio classifier*, which we found to be inaccurate and difficult to train, thereby circumventing the limitations faced in (Choi et al. 2020).

Next, we take it a step further, and consider an ***alternative, challenging problem setup***. In this setup, only pre-trained (potentially biased) models are available, while datasets that were used to pre-train the models are *inaccessible*. We show that proposed fairTL and fairTL++ methods are also effective under this setup, where they improve both quality and fairness of a pre-trained GAN by adapting it on a small fair dataset. We remark that since previous work requires access to the large dataset D_{bias}, it is incapable of handling this challenging setup. The significance of this new setup is that it enables fair and high-quality GANs without imposing access to large datasets and high computational resources.

Our **main contributions** are:

- In the Choi *et al.* setup (which assumes availability of both D_{bias} and D_{ref}) we show that a simple transfer learning approach –called fairTL– is very effective for training a fair generative model. We have also proposed fairTL++ by introducing two simple improvements upon fairTL to preserve general knowledge while capturing the fair distribution *w.r.t.* SAs during adaptation.

- We also introduce a more challenging setup which considers debiasing pre-trained GANs, where only the small fair dataset D_{ref} is available. Both proposed fairTL and fairTL++ approaches remain effective in this setup, paving the way for making better use of pre-trained GANs while addressing fairness.

- We conduct extensive experiments to show that our proposed method can achieve state-of-the-art (SOTA) performance in generated samples quality, diversity and fairness.

Related Work

Fairness in Generative Models. Fairness in machine learning (ML) is mostly studied for classification problems, where generally the objective is to handle a classification task independent of a SA in the input data, *e.g.* making 'hiring' decisions independent of `Gender`. Different measurement metrics are used for this objective, including well-known Equalised Odds, Equalised Opportunity (Hardt, Price, and Srebro 2016) and Demographic Parity (Feldman et al. 2015). However, in generative models, fairness is defined as *equal representation*, *i.e.* uniform distribution of samples *w.r.t.* SAs. This results in some misalignment in the objective of fair generative models with earlier classifier works.

Several works have addressed the enforcement of fairness in generative models, often with the use of auxiliary models. Fair-GAN (Xu et al. 2018) and FairnessGAN (Sattigeri et al. 2019) are proposed to generate fair datasets (data-points and labels) as a pre-processing technique. In these works, a downstream classifier learns to identify the SA, providing feedback to the generator. Nonetheless, all of

these works are supervised and hence require a large, well-labeled dataset. However in the proposed setup, we do not have access to such a labeled dataset.

Regardless, there exists a few works that adopt a similar unsupervised or semi-supervised approach. In particular, Fair GAN without retraining (Tan, Shen, and Zhou 2020), aims to learn the latent distributions of the input noise *w.r.t.* the SA, which then allows us to sample uniformly from it. Frankle *et al.* (Frankel and Vendrow 2020) introduces the concept of prior modification, where an additional smaller network is added to modify the prior of a GAN to achieve a fairer output. Importance weighting algorithm is proposed in (Choi et al. 2020) for the training of a fair generative model. In this algorithm, a reference fair dataset *w.r.t.* the SA is used during training, while simultaneously exposing the model to the large biased dataset (from which samples are re-weighted). This allows the generator to output high-quality samples, while encouraging fairness *w.r.t.* the SA. SOTA quality and fairness of generated samples has been reported in (Choi et al. 2020). Lastly, although not explored deeply, MaGNET (Humayun, Balestriero, and Baraniuk 2022) hints at the possibility that enforcing uniformity in the latent feature space of a GAN through a sampling process, may have an impact in enforcing fairness *w.r.t.* a SA.

Transfer Learning. The main idea in transfer learning is to achieve a low generalization risk by adapting a pre-trained model (usually trained on a large-diverse dataset) to a target domain/task by using usually limited data from the same target domain/task (Pan and Yang 2009; Zhao et al. 2022; Cong et al. 2020; Zhao, Cong, and Carin 2020; Mo, Cho, and Shin 2020). Generally, in *discriminative learning*, the pre-trained model is adapted in two simple ways (Yosinski et al. 2014; Jiang et al. 2022): i) *Linear-Probing* (LP), which freezes the pre-trained network weights and trains the newly added ones (Wu, Zhang, and Ré 2020; Malekzadeh et al. 2017; Du et al. 2020), and ii) *fine-tuning* (FT) which continues to train using the entire pre-trained network weights (Cai et al. 2019; Guo et al. 2019; Abdollahzadeh, Malekzadeh, and Cheung 2021). Recently, (Kumar et al. 2022) suggests that utilizing Linear-Probing prior to Fine-Tuning (LP-FT) can help preserve important features needed for adaptation. In *generative learning*, TGAN (Wang et al. 2018b) demonstrates the effectiveness of transferring pre-trained GANs into new domains, thereby improving performance with limited data. CDC (Ojha et al. 2021) uses a similar approach in Few-shot Cross-domain Adaptation, but with the addition of a cross-domain consistency loss. EWC (Li et al. 2020) discusses the preservation of certain weights during adaptation to maintain the diversity of the source domain. In contrast to the previous works that aims to address the improvement of sample quality on the target domain, we address a different concept –*improving the fairness* using transfer learning.

Multiple Feedback Approach. Learning through a multiple feedback approach has been a popular approach in improving quality of generated samples, particularly when faced with limited samples (Kumari et al. 2022; Tran et al. 2021). Instead of the standard one-generator-and-discriminator approach, the multiple feedback approach takes advantage of multiple discriminators (Nguyen et al.

2017; Durugkar, Gemp, and Mahadevan 2017; Albuquerque et al. 2022; Um and Suh 2021) or multiple generators, thereby improving stability during optimization (Hoang et al. 2022; Ghosh et al. 2018).

Proposed Method

In this section, we first consider the problem setup in (Choi et al. 2020) – which assumes the availability of D_{bias} and D_{ref} – and outline the details of the proposed fairTL and its improved variant fairTL++. Next, we describe a new challenging problem setup that removes the need for a large biased dataset D_{bias}, and only considers the availability of a pre-trained (possibly biased) GAN and a small fair dataset D_{ref}. Existing methods can not handle this setup because of their reliance on D_{bias} for training a fair GAN.

fairTL

Here, we present a simple transfer learning-based method in training a GAN for fair, diverse and high-quality sample generation, based on D_{bias} and D_{ref}. This process includes a pre-training step, which is followed by adaptation. In the pre-training stage, we train the generative model to learn the required general knowledge for sample generation, using all available training data. In particular, we train the model with GAN loss (Goodfellow et al. 2014). We remark that our approach is not restricted to a particular loss function. Here, we define G_s and D_s as the biased generator and discriminator in the pre-training stage, trained on samples in $D_{bias} \cup D_{ref}$. Next, in the adaptation stage, using the same loss function, we adapt the generative model to learn fair SA distribution by using D_{ref} only:

$$
\begin{aligned}
\min_{G_t} \max_{D_t} \mathcal{L} = & \mathbb{E}_{x \in D_{ref}}[\log D_t(x)] \\
& + \mathbb{E}_{z \sim p_z(z)}[\log (1 - D_t(G_t(z)))].
\end{aligned}
\tag{1}
$$

Here, G_t, D_t are generators and discriminators in the adaptation stage, trained on samples *only* in D_{ref}, and z is random noise sampled from a Gaussian noise distribution $p_z(z)$. Furthermore, G_t, D_t are initialized from G_s, D_s respectively. Our experimental results show that this simple approach can be considered as a *strong baseline* for fair GAN training which achieves competitive performance with the SOTA method proposed in (Choi et al. 2020).

fairTL++

One technical challenge of using fairTL is that due to the small size of D_{ref}, fine-tuning on D_{ref} is susceptible to mode collapse (Mo, Cho, and Shin 2020; Li et al. 2020). To prevent the model from forgetting the general knowledge learned during pre-training, we propose *fairTL++*, which includes two additions when adapting to D_{ref}: Linear-Probing before Fine-Tuning (LP-FT), and a multiple feedback approach during adaptation (Figure. 1). In what follows, we discuss the details of each method.

a) LP-FT. (Kumar et al. 2022) demonstrates that when adapting a pre-trained classifier to a new task, it is advantageous to first use Linear-Probing (updating the classifier head but freezing lower layers) for some limited epochs

T, and then use Fine-Tuning (updating all model parameters). This method is termed LP-FT. Experimental results in (Kumar et al. 2022) suggests that Linear-Probing allows for more task-specific parameters to adapt before Fine-Tuning, and generally works better for transfer learning. We found that a similar approach can be adopted for our generative learning setup. In our context, the discriminator can be considered as the feature extractor, and the downstream task is to learn the fairer SA distribution of D_{ref}.

To implement this, we first conduct an empirical study to identify the SA-specific layers needed for adaptation. In this study, we similarly implement fairness adaptation with fairTL, but with a large D_{ref}, thereby alleviating the instability during training. Next, we evaluated the mean layer weight change. In our results, we observed that amongst the layers in D_t and G_t only the first two layers of D_t (closest to the model's input) expressed low changes in their weight, thereby indicating that they are the least associated with the SA. Hence, these are general layers that should be preserved. To validate this, we implemented LP while freezing different layer permutations and similarly found that freezing any additional layers, other than the first two layers of D_t, resulted in poorer sample quality. We found that these results were consistent across several different SA. This finding aligns with works from domain adaptation (Mo, Cho, and Shin 2020), which similarly found it advantageous to retain (freeze) the lower-level layers of the discriminator *throughout* fine-tuning. However, we noted that retaining the low-level layers throughout the adaptation stage creates instability *i.e.* the generator start to output noise. Conversely, retaining those same layers for only T epoch improves quality and fairness. Therefore, when adapting G_s and D_s into fair dataset D_{ref}, we first freeze the lower layers of the discriminator for some limited epochs, and then fine-tune all parameters.

b) Multiple Feedback. (Kumari et al. 2022; Um and Suh 2021) proposes that the utilization of collective knowledge from multiple pre-trained discriminators improves GAN performance under limited data settings. Inspired by this, we consider that our pre-trained discriminator D_s is proficient at evaluating the quality of our generated samples despite being trained on a biased dataset. With this, we adopted a multiple feedback approach during our adaptation stage. In particular, we retain a *frozen* copy of our discriminator D_s after pre-training and append it to our model. Similarly, we carry out adaptation on D_{ref} with D_s, D_t and G_t. During this process, only D_t and G_t weights are updated. Intuitively, D_s can be seen to discriminate upon the generated sample quality, while the D_t adapts to the D_{ref} and enforces the new fair SA distribution. Eqn. 2 presents the loss function, where we utilize $\lambda \in [0, 1]$ as a hyper-parameter to control the balance between enforcing fairness and quality. In our experiments, we found that although both discriminators play an essential part in improving the performance of the GAN, more emphasis should be placed on D_t. In particular, since D_s is frozen, making λ too small results in instability during training. Conversely, making λ too big limits the feedback we get on the sample's quality. Empirically, we found $\lambda = 0.6$ to be ideal.

Perc	90_10				Multi			
	0.25	0.1	0.05	0.025	0.25	0.1	0.05	0.025
a) Imp-weighting (Choi et al. 2020)								
FID	19.20 ± 0.10	20.42 ± 0.20	23.01 ± 0.15	25.82 ± 0.13	14.61 ± 0.21	16.92 ± 0.31	19.43 ± 0.23	22.80 ± 0.13
FD	0.090 ± 0.011	0.107 ± 0.022	0.167 ± 0.016	0.246 ± 0.032	0.142 ± 0.032	0.116 ± 0.020	0.135 ± 0.014	0.144 ± 0.016
b) fairTL								
FID	$\underline{14.21} \pm 0.02$	$\underline{20.00} \pm 0.10$	$\underline{22.99} \pm 0.09$	$\underline{23.60} \pm 0.11$	$\underline{11.98} \pm 0.12$	$\underline{13.10} \pm 0.14$	$\underline{13.29} \pm 0.16$	$\underline{13.99} \pm 0.10$
FD	$\underline{0.087} \pm 0.007$	$\underline{0.105} \pm 0.020$	$\underline{0.107} \pm 0.012$	$\underline{0.130} \pm 0.029$	$\underline{0.113} \pm 0.021$	$\underline{0.115} \pm 0.017$	$\underline{0.118} \pm 0.013$	$\underline{0.138} \pm 0.011$
c) fairTL++								
FID	$\mathbf{9.02 \pm 0.03}$	$\mathbf{10.69 \pm 0.11}$	$\mathbf{20.12 \pm 0.04}$	$\mathbf{20.70 \pm 0.08}$	$\mathbf{10.50 \pm 0.10}$	$\mathbf{11.38 \pm 0.11}$	$\mathbf{12.00 \pm 0.10}$	$\mathbf{13.18 \pm 0.06}$
FD	$\mathbf{0.010 \pm 0.007}$	$\mathbf{0.062 \pm 0.022}$	$\mathbf{0.035 \pm 0.034}$	$\mathbf{0.092 \pm 0.025}$	$\mathbf{0.016 \pm 0.010}$	$\mathbf{0.090 \pm 0.020}$	$\mathbf{0.086 \pm 0.020}$	$\mathbf{0.101 \pm 0.016}$

Table 1: Comparing our proposed Fair Transfer Learning against Imp-weighting (Choi et al. 2020) on CelebA (Liu et al. 2015), for single SA (Gender) and multi-SA ({Gender,Blackhair}). For *single SA* (Gender), we utilize a D_{bias} with $bias = 90\_10$ *i.e.* 90% sample are Female and 10% Male, and for *multi-SA* a D_{bias} with bias F-NBH, F-BH, M-NBH, M-BH=$[0.437, 0.063, 0.415, 0.085]$ (Male(M), Female(F), BlackHair(BH) and No-BlackHair(NBH)). Then, we varied the sample size of D_{ref}, while $|D_{bias}|$ is kept constant. This is denoted by the ratio $perc = |D_{ref}|/|D_{bias}|$ for $\{0.25, 0.1, 0.05, 0.025\}$. With this setup, we utilize BIGGAN (Brock, Donahue, and Simonyan 2019) to reproduce (a) (Choi et al. 2020) the current SOTA results, and implement our proposed (b) fairTL and (C) fairTL++. We show that our proposed method fairTL is effective in achieving new SOTA FID and FD results for all $perc$, while fairTL++ demonstrates even greater improvements. For FID ($\downarrow$) and FD ($\downarrow$) , a low score indicates higher quality samples and fairer SA distribution, respectively.

Perc	0.25	0.1
a) Imp-weighting (Choi et al. 2020)		
FID($\downarrow$)	27.57 ± 0.45	39.03 ± 0.72
FD ($\downarrow$)	0.154 ± 0.031	0.205 ± 0.044
b) fairTL		
FID($\downarrow$)	$\underline{20.70} \pm 0.32$	$\underline{22.92} \pm 0.22$
FD ($\downarrow$)	$\underline{0.044} \pm 0.017$	$\underline{0.039} \pm 0.015$
c) fairTL++		
FID($\downarrow$)	$\mathbf{19.21 \pm 0.32}$	$\mathbf{21.22 \pm 0.19}$
FD ($\downarrow$)	$\mathbf{0.018 \pm 0.020}$	$\mathbf{0.003 \pm 0.002}$

Table 2: Comparing our proposed Fair Transfer Learning against Imp-weighting (Choi et al. 2020) on UTK-Face (Zhang, Song, and Qi 2017), for *single SA* (Race-Caucasian) . We utilize the same *single SA* setup, as per Tab. 1. Given that UTKFace is a small dataset, we are limited to $perc = \{0.25, 0.1\}$. We then similarly utilize BIG-GAN and compare a) (Choi et al. 2020) the current SOTA against our proposed b) fairTL and c) fairTL++. Similarly, we find that our proposed solutions outperform Choi *et al.* in both FID and FD.

$$\min_{G_t} \max_{D_t} \mathcal{L} = \mathbb{E}_{x \in D_{ref}}[\log D_t(x)]$$
$$+ \lambda \mathbb{E}_{z \sim p_z(z)}[\log (1 - D_t(G_t(z)))] \qquad (2)$$
$$+ (1 - \lambda) \mathbb{E}_{z \sim p_z(z)}[\log (1 - D_s(G_t(z)))]$$

As we will later discuss in our experiments, having the addition of LP-FT and multiple feedback approach improves the stability of our training process, and allows our proposed method to achieve SOTA quality and fairness.

Improving the Fairness of Pre-trained GANs

As mentioned before, when discussing our proposed fairTL and fairTL++, we assume that similar to (Choi et al. 2020), we have access to a large biased dataset D_{bias}, and a small fair dataset D_{fair}. Under this setup, the generative model requires to be trained from scratch, which entails significant computational resources. Also, a large dataset is necessary to have expressive representation. Another solution to learn a fair generative model, which can output diverse and high-quality samples, is to take advantage of (potentially biased) pre-trained generative models, and improve their fairness *w.r.t.* the desired SAs. Under this new challenging setup, we assume that there is a pre-trained GAN, but the dataset used for pre-training is *inaccessible*. However, we have access to a small, fair dataset from the related distribution. Since our proposed fairTL and fairTL++ methods are based on the general idea of transfer learning, they can be easily adapted to this challenging setup by discarding the first pre-training step. Our experimental results show that fairTL and fairTL++ remain effective under this setup, and can improve the fairness and quality of SOTA pre-trained GANs.

Experiments

In this section, we evaluate the performance of the proposed fairTL and fairTL++ in two different problem setups: 1) problem setup of (Choi et al. 2020) where both D_{bias} and D_{ref} are available for a given SA, 2) the proposed problem setup in this work where we have access to only the small D_{ref} and a pre-trained GAN in-place of D_{bias}.

Perc	0.025				
Sensitive Attributes	Gender	BlackHair	Young	Smiling	Moustache
a) Pre-trained					
FID($\downarrow$)	9.20 ± 0.02	14.58 ± 0.11	24.60 ± 0.21	9.30 ± 0.03	19.84 ± 0.21
FD ($\downarrow$)	0.102 ± 0.019	0.075 ± 0.002	0.277 ± 0.012	0.168 ± 0.007	0.376 ± 0.041
b) fairTL					
FID($\downarrow$)	9.01 ± 0.01	13.39 ± 0.09	12.94 ± 0.10	9.15 ± 0.01	13.03 ± 0.14
FD ($\downarrow$)	0.088 ± 0.010	0.058 ± 0.012	0.093 ± 0.023	0.098 ± 0.020	0.096 ± 0.042
c) fairTL++					
FID ($\downarrow$)	$\mathbf{8.81 \pm 0.01}$	$\mathbf{12.32 \pm 0.10}$	$\mathbf{11.79 \pm 0.12}$	$\mathbf{8.90 \pm 0.02}$	$\mathbf{11.66 \pm 0.14}$
FD ($\downarrow$)	$\mathbf{0.067 \pm 0.014}$	$\mathbf{0.057 \pm 0.008}$	$\mathbf{0.056 \pm 0.011}$	$\mathbf{0.061 \pm 0.023}$	$\mathbf{0.025 \pm 0.028}$

Table 3: Evaluating our proposed Fair Transfer Learning on a pre-trained StyleGAN2 (Karras et al. 2020) and FFHQ (Karras, Laine, and Aila 2019b) dataset. We evaluate our proposed method on the SA {Gender, Blackhair, Young, Smiling, Moustache}. As our baseline, we first evaluate the a) Pre-trained model's FID and Fairness (FD) *w.r.t.* the different SA. Then, utilising a $perc = |D_{ref}|/|D_{bias}|$ of 0.025 , we implement b) fairTL and c) fairTL++ and similarly measure the debiased StyleGAN2 FID and FD. Based on our results, we demonstrate that our proposed method is advantageous across SA in improving diversity, quality and fairness of the generated samples *w.r.t.* the SA.

For the first setup, we compare our proposed method against importance weighting (Choi et al. 2020) which produces SOTA in quality and fairness. As importance weighting (Choi et al. 2020) cannot be applied to the second setup due to the unavailability of the large dataset D_{bias}, we evaluate the performance of the proposed method on mitigating the (potential) existing bias in SOTA pre-trained GANs. We remark that in both setups, *none of the fairness enforcement methods have access to the labels of the datasets* and that these labels are only used as a controlled means to re-sample the respective datasets and simulate the bias.

Evaluation Metric. Following (Choi et al. 2020), we utilize FID (Heusel et al. 2018) to evaluate the quality and diversity of our generated samples, and the fairness discrepancy metric (FD) (Choi et al. 2020) to measure the fairness of our models *w.r.t.* a SA. Similar to (Choi et al. 2020), when evaluating FID, we re-sample the original large dataset (*e.g.* CelebA) to obtain equal SA representation, which we use to calculate the reference statistics. This is necessary as it allows us to obtain an estimate of the quality and diversity of the generator, while referencing our target ideal generator with fair SA distribution. Then, to evaluate fairness, we train a ResNet-18 (He et al. 2016) to classify the generated samples SA, which we use to calculate FD as follows:

$$f = |\bar{p} - \mathbb{E}_{z \sim p_z(z)}[C(\mathbf{G}(\mathbf{z}))]|_2 \qquad (3)$$

Here, $C(\mathbf{G}(z))$ is the one-hot vector for the classified label of the generated sample $\mathbf{G}(z)$, whose generator can either be $\mathbf{G}_t$ or $\mathbf{G}_s$ depending on the method used. z is sampled from a Gaussian noise distribution $p_z(z)$ and $\bar{p}$ is a uniformly distributed vector with the same cardinality as $C(\mathbf{G}(\mathbf{z}))$.

Setup 1: Training a Fair Generator

Utilising the setup from (Choi et al. 2020), we implement our proposed method, by first training a BIGGAN (Brock, Donahue, and Simonyan 2019) model with all the available data ($D_{bias} \cup D_{ref}$) to achieve the highest quality generator. This is then followed by the adaptation stage, with D_{ref} only. For a fair comparison, we utilize the source code from (Choi et al. 2020) to reproduce their proposed importance-weighting (imp-weighting) on BIGGAN.

Dataset. We consider the datasets CelebA (Liu et al. 2015) and UTKFace (Zhang, Song, and Qi 2017) for this experiment. For CelebA, following Choi *et al.*, we utilize the SA Gender and {Gender,Blackhair} for both single and multi-attribute settings respectively. Then, for UTKFace, we utilize the SA Race(Caucasian). In both single attribute settings, we synthetically introduce a $bias = 0.9$ to D_{bias} *i.e.* D_{bias} contains 90% Female/Caucasian samples and 10% Male/Non-Caucasian samples, by re-sampling the dataset. For multi-attribute settings, given the data limitations, we similarly introduce a D_{bias} through re-sampling with the following sample ratios F-NBH, F-BH, M-NBH, M-BH= $[0.437, 0.063, 0.415, 0.085]$ for Male(M), Female(F), Blackhair(BH) and no-Blackhair(NBH). Next, we considered different $|D_{ref}|$ while keeping $|D_{bias}|$ constant, notated by $perc = |D_{ref}|/|D_{bias}|$. This allows us to evaluate the robustness of our proposed method with decreasing reference samples during adaptation. For the CelebA dataset, we explore $perc = \{0.25, 0.1, 0.05, 0.025\}$ and for UTKFace, due to its smaller dataset, we explore $perc = \{0.25, 0.1\}$.

Single Attribute Results. Table 1 presents the results of imp-weighting (Choi et al. 2020) against our proposed methods on the CelebA dataset. Comparing across different $perc$, we observe that fairTL is generally able to outperform imp-weighting, achieving better fairness and quality. Then, with the addition of LP-FT and the multi-feedback approach, we observed greater improvements in fairTL++, highlighting the effectiveness of these two additions during adaptation. In particular, we notice that even with the smallest reference dataset, $perc = 0.025$, fairTL++ is able to achieve a relatively fair generator, *i.e.* low FD measurement, while
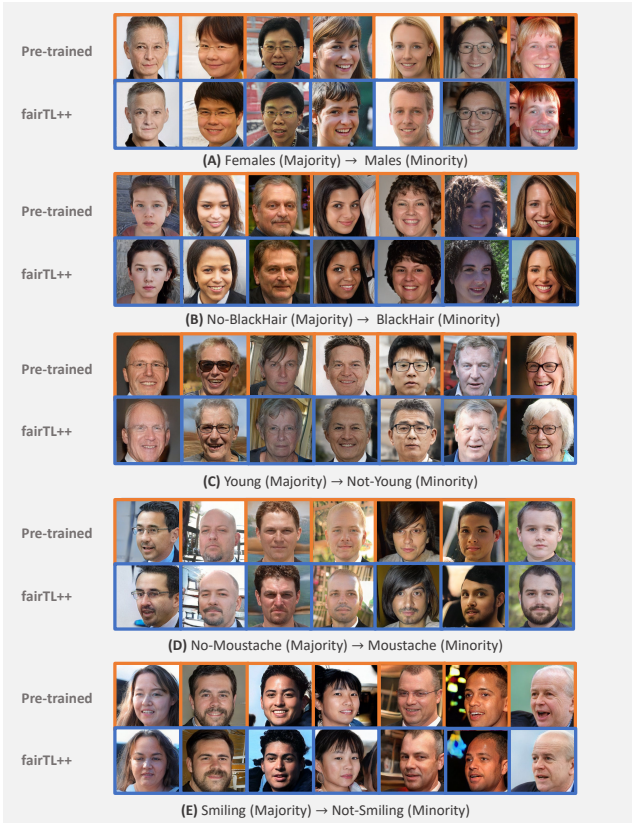
Figure 2: Illustration of samples before and after fairness adaptation by our fairTL++ on a pre-trained Style-GAN2 (Karras et al. 2020). For each sample, we utilize the same noise vector to sample from the pre-trained model and fairTL++ after SA adaptation. Notice how the samples are adapted from the majority to minority represented SA.

imp-weighting worsens under these conditions. Table 2 then compares the same methods but on the UTKFace dataset with SA Race-Caucasian. With this dataset, we similarly observe that fairTL++ outperforms both imp-weighting and fairTL in quality and fairness. In fact, on the smaller UTK-Face dataset, the benefits of our proposed method becomes more prominent with imp-weighting's sample quality (FID) significantly degrading as $perc$ becomes smaller. In contrast, our proposed method only experiences minor degradation, while still enforcing SOTA fairness (FD).

Multi-attribute results. Table 1 presents a similar experiment but with multiple SA Gender and Blackhair. Our results show that even under this more challenging setup, involving two SA simultaneously, fairTL++ still outperforms imp-weighting, thereby achieving SOTA performance in mitigating bias while maintaining high-quality samples.

Setup 2: Debiasing a Pre-trained Generator

In this new setup, we demonstrate that unlike previous works, our proposed method does not strictly require access to the large dataset (D_{bias}). Instead, we are able to improve on the fairness of existing biased pre-trained models.

For this experiment, we utilize the original code as in Style-GAN2 (Karras et al. 2020) as the baseline, along with the pre-trained weights on the FFHQ dataset (Karras, Laine, and Aila 2019b). With this baseline, we followed the same setup as the previous experiments for fair comparison, and measured the FID and FD of the pre-trained model across different SA. Then utilizing D_{ref} we implement the adaptation stage for fairTL and fairTL++ and re-evaluated the model.

Dataset. We utilize the FFHQ dataset and consider the SA {Gender, Blackhair, Young, Smiling, Moustache} to demonstrate the effectiveness of our proposed method across different SA. For each SA, we attained a D_{ref} with $perc$=0.025.

From our results in Table 3, we observe that the pre-trained StyleGAN2 model contains a considerable amount of bias in the selected SA. In particular, larger biases exist for SA {Young,Smiling,Moustache} , where high FD measurements were reported. Furthermore, the high FID measurements indicates a mismatch between the the diversity of the generated samples and the ideal reference samples. Our proposed solutions however proves to be effective in improving both fairness and diversity of the StyleGAN2, while achieving high-quality samples, as seen from the relatively low FD and FID score. Similar to the previous experiments, fairTL++ proves to be the more effective method. Fig. 2 illustrates a few samples that have been adapted from the majority-represented SA to the minority-represented SA, thereby achieving a fairer SA distribution. We remark that though the SA of the samples have been adapted *e.g.* Female to Male, the underlying general attribute *e.g.* pose and race remain unchanged.

Conclusion

In our work, we focus on the challenging task of training a diverse, high-quality GAN while achieving fairness *w.r.t.* some sensitive attributes. In this task, we are given the real world constraints of having only access to a small but fair dataset and a large but biased dataset. To overcome these limitations, we propose a simple and effective method of training a fair generative model via transfer learning. To do this, we first pre-train the model with the large biased dataset, followed by fairness adaptation with the small unbiased dataset. We then further demonstrate that the introduction of a multiple feedback approach and Linear-Probing to the sensitive attribute specific layers during adaptation, can help further improve both sample quality and fairness, thereby achieving state-of-the-art performance. Additionally, we demonstrate that our proposed methods can similarly improve the quality and fairness of SOTA pre-trained GANs.

Acknowledgements

References

Abdollahzadeh, M.; Malekzadeh, T.; and Cheung, N.-M. M. 2021. Revisit Multimodal Meta-Learning through the Lens of Multi-Task Learning. *Advances in Neural Information Processing Systems*, 34.

Albuquerque, I.; Monteiro, J.; Doan, T.; Considine, B.; Falk, T.; and Mitliagkas, I. 2022. Multi-Objective Training of Generative Adversarial Networks with Multiple Discriminators. *ICLR19*.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv:1809.11096 [cs, stat]*.

Cai, H.; Wang, T.; Wu, Z.; Wang, K.; Lin, J.; and Han, S. 2019. On-device image classification with proxyless neural architecture search and quantization-aware fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Chizzola, V.; Micheli, B.; and Vingelli. 2017. TARGET – Taking a Reflexive Approach to Gender Equality for Institutional Transformation. *STEM Gender Equality Congress Proceedings*, 1(1): 839–839.

Choi, K.; Grover, A.; Singh, T.; Shu, R.; and Ermon, S. 2020. Fair Generative Modeling via Weak Supervision. In *Proceedings of the 37th International Conference on Machine Learning*, 1887–1898. PMLR.

Cong, Y.; Zhao, M.; Li, J.; Wang, S.; and Carin, L. 2020. GAN Memory with No Forgetting. *arXiv:2006.07543 [cs]*.

Du, S. S.; Hu, W.; Kakade, S. M.; Lee, J. D.; and Lei, Q. 2020. Few-Shot Learning via Learning the Representation, Provably. In *International Conference on Learning Representations*.

Durugkar, I.; Gemp, I.; and Mahadevan, S. 2017. GENERATIVE MULTI-ADVERSARIAL NETWORKS. *ICLR 2017*, 14.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 259–268.

Frankel, E.; and Vendrow, E. 2020. Fair Generation Through Prior Modification. *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*.

Frid-Adar, M.; Diamant, I.; Klang, E.; Amitai, M.; Goldberger, J.; and Greenspan, H. 2018. GAN-based synthetic medical image augmentation for increased CNN performance in liver lesion classification. *Neurocomputing*, 321: 321–331.

Ghosh, A.; Kulharia, V.; Namboodiri, V.; Torr, P. H. S.; and Dokania, P. K. 2018. Multi-Agent Diverse Generative Adversarial Networks. arXiv:1704.02906.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guo, Y.; Shi, H.; Kumar, A.; Grauman, K.; Rosing, T.; and Feris, R. 2019. Spottune: transfer learning through adaptive fine-tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4805–4814.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv:1706.08500 [cs, stat]*.

Hoang, Q.; Nguyen, T. D.; Le, T.; and Phung, D. 2022. MGAN: Training Generative Adversarial Nets with Multiple Generators. In *International Conference on Learning Representations*.

Humayun, A. I.; Balestriero, R.; and Baraniuk, R. 2022. MaGNET: Uniform Sampling from Deep Generative Network Manifolds Without Retraining. In *International Conference on Learning Representations*.

Hutchinson, B.; and Mitchell, M. 2019. 50 Years of Test (Un)Fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 49–58. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-4503-6125-5.

Hwang, S. 2020. FairFaceGAN: Fairness-aware Facial Image-to-Image Translation. *BMVS20*, 14.

Jalan, H. J.; Maurya, G.; Corda, C.; Dsouza, S.; and Panchal, D. 2020. Suspect Face Generation. In *2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA)*, 73–78.

Jiang, J.; Shu, Y.; Wang, J.; and Long, M. 2022. Transferability in Deep Learning: A Survey. *arXiv preprint arXiv:2201.05867*.

Karras, T.; Laine, S.; and Aila, T. 2019a. A style-based generator architecture for generative adversarial networks. In *CVPR*.

Karras, T.; Laine, S.; and Aila, T. 2019b. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. *arXiv:1912.04958 [cs, eess, stat]*.

Katal, A.; Wazid, M.; and Goudar, R. H. 2013. Big Data: Issues, Challenges, Tools and Good Practices. In *2013 Sixth International Conference on Contemporary Computing (IC3)*, 404–409. Noida, India: IEEE. ISBN 978-1-4799-0192-0 978-1-4799-0190-6.

Kumar, A.; Raghunathan, A.; Jones, R.; Ma, T.; and Liang, P. 2022. Fine-Tuning Can Distort Pretrained Features and Underperform Out-of-Distribution. *ICLR 2022*.

Kumari, N.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2022. Ensembling Off-the-Shelf Models for GAN Training. *CVPR22*, 12.

Le Quy, T.; Roy, A.; Iosifidis, V.; Zhang, W.; and Ntoutsi, E. 2022. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, e1452.

Li, Y.; Zhang, R.; Lu, J. C.; and Shechtman, E. 2020. Few-Shot Image Generation with Elastic Weight Consolidation. In *Advances in Neural Information Processing Systems*, volume 33, 15885–15896. Curran Associates, Inc.

Lim, S. K.; Loo, Y.; Tran, N.-T.; Cheung, N.-M.; Roig, G.; and Elovici, Y. 2018. DOPING: Generative Data Augmentation for Unsupervised Anomaly Detection. In *Proceeding of IEEE International Conference on Data Mining (ICDM)*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Lucas, A.; Lopez-Tapia, S.; Molina, R.; and Katsaggelos, A. K. 2019. Generative adversarial networks and perceptual losses for video super-resolution. *IEEE Transactions on Image Processing*, 28(7): 3312–3327.

Malekzadeh, T.; Abdollahzadeh, M.; Nejati, H.; and Cheung, N.-M. 2017. Aircraft fuselage defect detection using deep neural networks. *arXiv preprint arXiv:1712.09213*.

Mo, S.; Cho, M.; and Shin, J. 2020. Freeze the Discriminator: A Simple Baseline for Fine-Tuning GANs. *arXiv:2002.10964 [cs, stat]*.

Muehlethaler, C.; and Albert, R. 2021. Collecting data on textiles from the internet using web crawling and web scraping tools. *Forensic Science International*, 322: 110753.

Nasrollahi, H.; Farajzadeh, K.; Hosseini, V.; Zarezadeh, E.; and Abdollahzadeh, M. 2020. Deep artifact-free residual network for single-image super-resolution. *Signal, Image and Video Processing*, 14(2): 407–415.

Nguyen, T.; Le, T.; Vu, H.; and Phung, D. 2017. Dual Discriminator Generative Adversarial Nets. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 2667–2677. Curran Associates, Inc.

Ojha, U.; Li, Y.; Lu, J.; Efros, A. A.; Jae Lee, Y.; Shechtman, E.; and Zhang, R. 2021. Few-Shot Image Generation via Cross-domain Correspondence. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10738–10747. Nashville, TN, USA: IEEE. ISBN 978-1-66544-509-2.

Pan, S. J.; and Yang, Q. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10): 1345–1359.

Sattigeri, P.; Hoffman, S. C.; Chenthamarakshan, V.; and Varshney, K. R. 2019. Fairness GAN: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5): 3–1.

Schlegl, T.; Seeböck, P.; Waldstein, S. M.; Schmidt-Erfurth, U.; and Langs, G. 2017. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. *CoRR*, abs/1703.05921.

Tan, S.; Shen, Y.; and Zhou, B. 2020. Improving the Fairness of Deep Generative Models without Retraining. *arXiv:2012.04842 [cs]*.

Tran, N.-T.; Tran, V.-H.; Nguyen, N.-B.; Nguyen, T.-K.; and Cheung, N.-M. 2021. On Data Augmentation for GAN Training. *IEEE Transactions on Image Processing*, 30: 1882–1897.

Um, S.; and Suh, C. 2021. A Fair Generative Model Using Total Variation Distance. *OpenReview*.

Wang, C.; Xu, C.; Wang, C.; and Tao, D. 2018a. Perceptual adversarial networks for image-to-image transformation. *IEEE Transactions on Image Processing*, 27(8): 4066–4079.

Wang, Y.; Wu, C.; Herranz, L.; van de Weijer, J.; Gonzalez-Garcia, A.; and Raducanu, B. 2018b. Transferring GANs: Generating Images from Limited Data. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision – ECCV 2018*, volume 11210, 220–236. Cham: Springer International Publishing. ISBN 978-3-030-01230-4 978-3-030-01231-1.

Wu, S.; Zhang, H. R.; and Ré, C. 2020. Understanding and improving information transfer in multi-task learning. *arXiv preprint arXiv:2005.00944*.

Xu, D.; Yuan, S.; Zhang, L.; and Wu, X. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, 570–575. IEEE.

Yosinski, J.; Clune, J.; Bengio, Y.; and Lipson, H. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

Yu, B.; Zhou, L.; Wang, L.; Shi, Y.; Fripp, J.; and Bourgeat, P. 2019. Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis. *IEEE transactions on medical imaging*, 38(7): 1750–1762.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *CVPR*.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5810–5818.

Zhao, M.; Cong, Y.; and Carin, L. 2020. On Leveraging Pretrained GANs for Generation with Limited Data. In *Proceedings of the 37th International Conference on Machine Learning*, 11340–11351. PMLR.

Zhao, Y.; Chandrasegaran, K.; Abdollahzadeh, M.; and man Cheung, N. 2022. Few-shot Image Generation via Adaptation-Aware Kernel Modulation. In *Advances in Neural Information Processing Systems*.