

FAIRTL: A Transfer Learning Approach for Bias Mitigation in Deep Generative Models

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

Abstract—This work studies fair generative models. We reveal and quantify the biases in state-of-the-art (SOTA) GANs *w.r.t.* different sensitive attributes. To address the biases, our main contribution is to propose novel methods to learn fair generative models via transfer learning. Specifically, first, we propose FAIRTL where we pre-train the generative model with a large biased dataset, then adapt the model using a small fair reference dataset. Second, to further improve sample diversity, we propose FAIRTL++, containing two additional innovations: i) *aligned feature adaptation*, which preserves learned general knowledge while improving fairness by adapting only sensitive attribute-specific parameters, ii) *multiple feedback discrimination*, which introduces a frozen discriminator for quality feedback and another evolving discriminator for fairness feedback. Taking one step further, we consider an alternative challenging and practical setup. Here, only a pre-trained model is available but the dataset used to pre-train the model is inaccessible. We remark that previous work requires access to large, biased datasets and cannot handle this setup. Extensive experimental results show that FAIRTL and FAIRTL++ achieve state-of-the-art performance in quality, diversity and fairness in both setups. Code is available at: <https://github.com/sutd-visual-computing-group/FairTL>.

Index Terms—Fairness, Generative Models, Transfer Learning, Generative Adversarial Networks.

I. INTRODUCTION

DEEP generative models such as Generative Adversarial Networks (GANs) are an active research area [1]–[8] that have achieved outstanding results in many tasks *e.g.*, image and audio restoration [9]–[11], data augmentation [12], undetectable image manipulation [13], image synthesis [4], [14], 2D to 3D image synthesis [15] image transformation [16], [17], super-resolution [8], text-to-image synthesis [18], [19], video captioning [7], image fusion [20], [21] and anomaly detection [22]. Given these advancements, generative models have been increasingly adopted in various real-world applications including high-stakes areas such as criminal justice [23], healthcare [24] and even supplementation of training data for down-stream tasks [25].

With these sensitive applications, increasing concerns have been raised regarding the efficacy of these generative models. This is mainly due to the potential biases that may exist in popular generative models, which may then perpetuate to downstream tasks and result in possible societal impact. For example, in suspect facial profiling [23] a generative model with Gender or Race biases, could result in wrongful

incrimination of an individual. As another example, when supplementing training data for downstream tasks [24], bias in the training dataset’s distributions can result in degraded performance *e.g.*, classifiers for disease diagnosis experience performance degradation when trained with datasets with class imbalanced *w.r.t.* Gender [26].

To address these concerns, **fairness in generative models** has recently received a lot of attention [27]–[30]. In these works, fairness is defined as *equal representation* [31] of some selected *Sensitive Attribute* (SA). For example, a generative model with an equal probability of producing Male and Female samples is fair *w.r.t.* Gender.

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. Ideally, these training datasets are large and assumed to be unbiased *w.r.t.* SAs. However, the unbiased condition usually holds only when we follow good practices for data collection *e.g.*, protocols adopted by governmental and international institutions such as the World Bank [28], [33]. Practically, these protocols are usually unscalable and collected *fair datasets are usually small* [28]. Therefore, to collect the required large dataset, we usually use alternative sources such as scraping images from the internet [34]. Collected data from these *alternative resources are usually biased w.r.t.* SAs [35], [36], and these biases are easily picked up by the generative models. As a result, dataset biases are one of the major causes of biases in deep generative models [37].

To prevent the biased dataset from harming the fairness of the generative model, the large biased dataset $\mathcal{D}_{bias}$ can be augmented with a small fair (*w.r.t.* some specific SAs) dataset $\mathcal{D}_{ref}$ [28]. In this setup, the main idea is that the generative model can learn expressive representation using $\mathcal{D}_{bias}$ while mitigating the bias with $\mathcal{D}_{ref}$. We remark that neither datasets is labeled *w.r.t.* SAs, and the size of the $\mathcal{D}_{ref}$ can be much smaller than $\mathcal{D}_{bias}$, as the fair dataset is obtained via rigorous protocols. For example, $|\mathcal{D}_{ref}|$ could be 2.5% of $|\mathcal{D}_{bias}|$.

In our work, we propose new methods for training fair generative models via transfer learning (Fig. 1). To do this, first, we study the existing bias in current State-Of-The-Art (SOTA) GANs. To the best of our knowledge, we are the first to explicitly reveal and quantify that even when large and diverse datasets are used to train current SOTA GANs, the degree of bias *w.r.t.* various SAs is noticeable (Fig. 2). Then, for training a fair generative model, based on the setup in the existing literature [28] (Setup 1), we propose a simple and effective *transfer learning* approach for bias mitigation. Specifically, we first propose fair transfer learning

Equal Contribution [†]Corresponding Author

Christopher T.H.Teo, Milad Abdollahzadeh, and Ngai-Man Cheung are with the Singapore University of Technology and Design, 8 Somapah Rd, Singapore 487372 (email: christopher_teo@mymail.sutd.edu.sg, milad_abdollahzadeh@sutd.edu.sg, ngaiman_cheung@sutd.edu.sg)

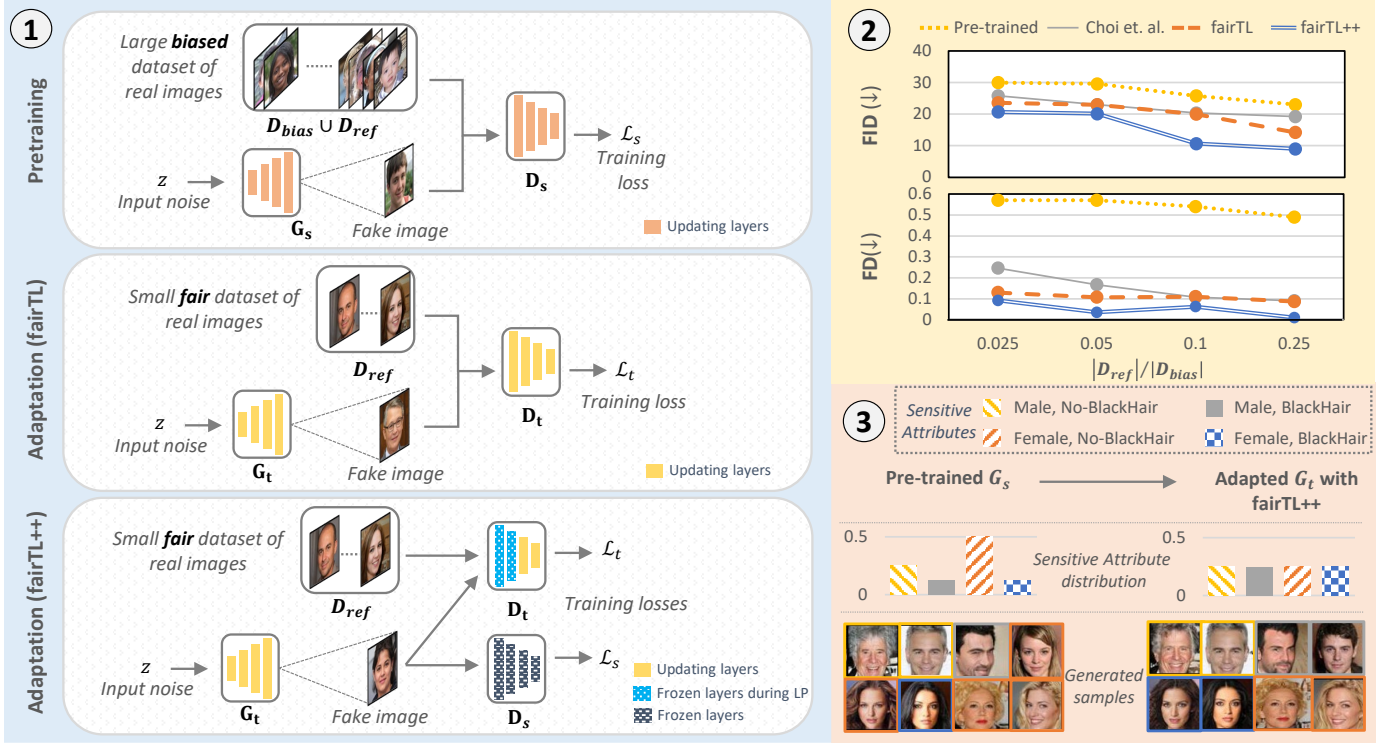


Fig. 1. **Overview of our work on training fair generative models.** ① We train a high-quality generator with fair sensitive attribute (SA) distribution in a two-stage process that we call FAIRTL. In *pre-training*, the GAN learns to generate diverse and high-quality samples from a large but biased dataset, D_{bias} . Then, in *adaptation*, the same GAN learns the fair underlying SA distribution from a small reference distribution, D_{ref} . To improve adaptation, we introduce an enhanced variation called FAIRTL++, which includes an additional source of feedback (D_s) and a Linear-Probing step before Fine-Tuning. The pre-training step is the 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 [32]. We compare four approaches 1) *pre-training*, 2) [28]: 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 a smaller FD indicates better fairness, see Eq. 5. Without consideration of fairness, the *pre-trained* setup expresses a large bias. Choi et al. [28] then significantly improves on this but brings about diminishing quality and fairness as $|D_{ref}|$ becomes smaller. Meanwhile, our proposed method demonstrates improved robustness, achieving SOTA results in both FID and FD. ③ We illustrate the improved fairness on multiple SA {Gender, BlackHair} during the adaptation stage. In these examples, we utilize a fixed noise vector, z to sample from both the pre-trained and FAIRTL++ models. We can observe how the majority-represented SA is adapted to the minority-represented SA, thereby improving the SA distribution and mitigating the biases.

(FAIRTL) where we pre-train the generative model using all available datasets to learn expressive sample generation. Subsequently, on top of the learned expressive knowledge, we adapt the model using only the small fair dataset to capture the fair SA distribution. We show that this simple transfer learning approach can achieve a *strong baseline* for training a fair generative model (Figure 1). To further improve the quality, diversity, and fairness of the generated samples, we propose **FAIRTL++** where we include two additional improvements upon FAIRTL: i) aligned feature adaptation, where we identify the SA-specific parameters and use a two-step process to preserve general knowledge while adapting SA-specific knowledge into the fair dataset, ii) multiple feedback discrimination for adversarial training, where we introduce a frozen discriminator to provide constant quality feedback and an evolving discriminator to provide fairness feedback during adaptation. We find that these two innovations can achieve considerable gains when applied to FAIRTL, achieving SOTA performance in sample quality, diversity, and fairness. In our analysis, we show that FAIRTL and FAIRTL++ differentiate themselves by *departing from the use of density ratio classifier*, which we found to be inaccurate and difficult to train (details

in Sec. V-C), thereby circumventing the limitations faced in the existing SOTA [28].

Next, we take it a step further and consider an *alternative, challenging problem setup* (Setup 2). In this setup, only a pre-trained (potentially biased) generative model and the small fair dataset are made available, while the larger dataset used to pre-train the models is *inaccessible*. We show that the proposed FAIRTL and FAIRTL++ methods are also effective under this setup. Specifically, they improve both quality and fairness, of a pre-trained GAN, while preserving diversity, by adapting it to a small fair dataset. We remark that since the existing SOTA method [28] requires access to the large dataset D_{bias} for fair training, it is incapable of handling this more 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:

- Assuming the availability of a large biased dataset D_{bias} , and a small fair dataset D_{ref} (Setup 1), we show that pre-training generative model on the available data, and transfer learning on D_{ref} is a strong baseline for training fair generative models, coined as FAIRTL.

- We improve FAIRTL by introducing two additional modules: a) *Aligned Feature Adaptation* which detects and adapts the SA-specific layers, and b) *Balancing Quality and Fairness Objective* which introduces an additional quality feedback loop to the generator, resulting in FairTL++.
- We analyze the bias in current SOTA generators and show that despite being trained on large and diverse datasets, the generated samples show a considerable amount of bias.
- Motivated by our analysis, we introduce a new more challenging setup (Setup 2) to debias pre-trained generators. In this setup, only the pre-trained generator and a small fair dataset $\mathcal{D}_{ref}$ are accessible, but the data used for training the generator is unknown.
- Extensive experiments demonstrate that our approach is extremely effective, achieving SOTA performance in both setups.

This work is a significant extension of our conference work [38]. In particular, we include several new studies. Firstly, we analyze biases in existing SOTA GANs quantitatively. We remark that even though previous works have hinted at biases in GANs, to the best of our knowledge, none have rigorously studied and quantified them. This lack of understanding often results in mismatch between their experiment setups and biases in SOTA GANs, which we address in this work. Secondly, to gain better insight for FAIRTL++, we conduct an in-depth analysis to identify parameters that encode general knowledge and SA-specific knowledge. Informed by our analysis, we develop a technique to improve feature alignment during fairness adaptation. Thirdly, we perform in-depth analysis, comparing our work against the existing SOTA work [28]. Our new analysis in this work further reveals that the central issue in the existing SOTA approach is the inaccuracy of the density ratio classifier [28]. Our analysis supports our experimental results, as we address and eliminate this limitation from the previous work. Overall our extended studies and experiments provide insights on the contributions made by the different modules of our proposed methods.

The paper is organized as follows. In Sec. II we discuss the related work for fair generative models. Sec. III analyzes the bias in current SOTA GANs. In, Sec. IV the details of the proposed method are discussed. The experimental results are discussed in Sec. V, and the limitations and conclusion are made in Sec. VI and VII, respectively.

II. RELATED WORK

Fairness in Generative Models. Fairness in machine learning is mostly studied for classification problems, where generally the objective is to handle a classification task independent of an SA in the input data, *e.g.*, making ‘hiring’ decisions independent of Gender. Different measurement metrics are used for this objective *e.g.*, Equalized Odds, Equalized Opportunity [39] and Demographic Parity [40]. However, in generative models, fairness has been generally studied through the lens of *equal representation* [27], [28], [30], *i.e.*, uniform SA distribution.

Several works have addressed the enforcement of fairness in generative models. Fair-GAN [41] and FairnessGAN [42]

propose the use of auxiliary models to generate fair datasets as a pre-processing technique. However, these methods pose a major limitation in requiring labels (supervised). Meanwhile, there exists a few works that adopt an unsupervised or semi-supervised approach to improve fairness. FairGAN without retraining [30] learns the latent distributions of the input noise *w.r.t.* the SA, allowing for fair sampling. Frankle *et al.* [27] introduces a smaller network to modify the prior of the GAN. Importance weighting (Imp-weighting) [28] achieves SOTA fairness and image quality through a re-weighting algorithm allowing the GAN to learn from both a small fair dataset *w.r.t.* the SA, while still exposing the GAN to the large but biased dataset. Lastly, although not explored deeply, MaGNET [29] hints at the possibility that enforcing uniformity in the latent feature space 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 to a target domain/task by using usually limited data from the same target domain/task [43]–[46]. Generally, in *discriminative learning*, the pre-trained model is adapted in two simple ways [47], [48]: i) *Linear-Probing* (LP), which freezes the pre-trained network weights and trains the newly added ones [49]–[51], and ii) *fine-tuning* (FT) which continues to train using the entire pre-trained network weights [52]–[54]. Recently, [55] suggests that utilizing Linear-Probing before Fine-Tuning (LP-FT) can help preserve important features needed for adaptation. In *generative learning*, TGAN [56] demonstrates the effectiveness of transferring pre-trained GANs into new domains with limited data. Recently, several works are proposed on top of TGAN to preserve the knowledge of the pre-trained GAN during adaptation: EWC [57] penalizes the update on important weights of the GAN by computing fisher information on each weight, CDC [6] introduces cross-domain consistency loss, DCL [58] uses mutual information between generated samples by pre-trained and adapted GANs, and AdAM [43] identifies the important kernels for adapting from a source to a target domain and preserves them by low-rank modulation. In contrast to the previous works that aim 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 shown to improve sample quality under limited data [59], [60]. Instead of the standard one-generator-and-discriminator approach, the multiple-feedback approach utilizes multiple discriminators [61]–[64] or generators, to improve optimization stability [65], [66].

III. ANALYZING BIAS OF SOTA GANs

SOTA GANs are usually trained on large datasets, *e.g.*, StyleGAN2 is trained on the FFHQ dataset [68], and ProGAN is trained on the CelebA-HQ dataset [67]. After training, these GANs can generate diverse and high-quality images that have a similar distribution to these datasets (measured by FID [69]). However, there is a lack of studies to understand if this high diversity would also result in fair GAN *w.r.t.* different SAs like

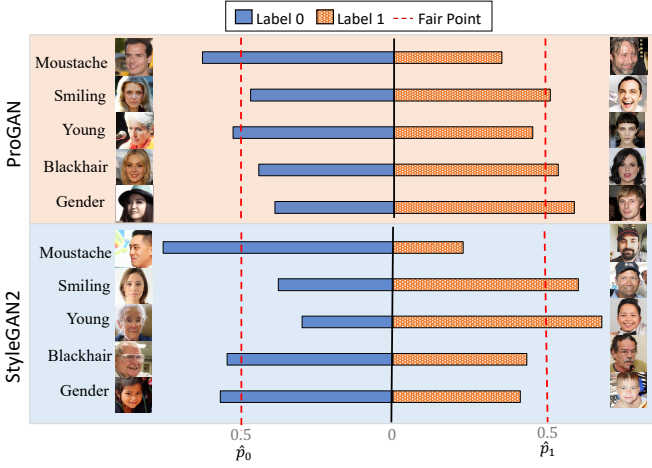


Fig. 2. Our study to quantify biases in SOTA ProGAN [4] trained on CelebA-HQ dataset [67], and StyleGAN2 [5] trained on FFHQ [68]. Utilizing auxiliary classifiers we measure the SA distribution, $\hat{p}=[\hat{p}_0, \hat{p}_1]$ for both ProGAN and StyleGAN2 *w.r.t.* SA {Gender(Male), BlackHair, Young, Moustache, Smiling}. The red-dotted line indicates an ideal fair (even) SA distribution.

Gender, BlackHair, etc. To gain insights into the fairness of existing GANs, in this section, we perform an analysis to quantify biases in the generated images of SOTA GANs *w.r.t.* various SAs.

Bias Analysis Setup. We utilize prominent, pre-trained StyleGAN2¹ [5] and ProGAN² [4] on FFHQ [68] and CelebA-HQ [4] datasets, respectively. Then to study bias, we generated 10K samples from each GAN. However, as the generated samples are unlabelled, we apply SA classifiers to classify the generated samples *w.r.t.* the desired SA. More specifically, for each GAN, and for each SA, we train a separate ResNet-18 SA classifier on the respective dataset. We use the following SA for each GAN in this study: {Moustache, Smiling, Young, BlackHair, Gender}. The SA classifiers for ProGAN are trained on the CelebA-HQ dataset where the labels for our desired SAs in this experiment are available [32]. Similarly, we use the FFHQ dataset to train the SA classifiers for StyleGAN2. Note that since the FFHQ dataset is not accompanied by labels [68], we follow [70] to attain publicly available³ soft labels *w.r.t.* desired SAs. These soft labels are then converted into hard labels by simply taking the $\arg \max$, and used for training the SA classifiers. Finally, we remark that when training each SA classifier, we make the related dataset fair *w.r.t.* the SA, thereby preventing the possibility of bias leakages to the trained SA classifiers. For example, when training a Gender classifier for ProGAN, as training data, we use the largest available subset of CelebA-HQ that has the same number of Male and Female samples. Overall our SA classifiers achieved high average accuracy in the range of [90%, 97%]. After training the SA classifiers, we pass the generated 10K samples by the GANs to each of the classifiers, and measure the SA distribution as follows: $\hat{p} = \mathbb{E}_{z \sim p_z}[C(G(z))]$. Here $C(G(z))$ is the one-hot vector

for the classified label of the generated sample $G(z)$, and z is sampled from a Gaussian noise distribution p_z .

Based on the above analysis, we obtain Fig. 2, the SA distribution ($\hat{p}$) for attributes: {Moustache, Smiling, Young, BlackHair, Gender}. From Fig. 2, we observe:

- **Observation 1:** Comparing results with an expected fair image generation (dashed red lines with $\hat{p}_0 = \hat{p}_1 = 0.5$), we can see that current SOTA GANs are biased *w.r.t.* the evaluated SAs even though they are trained on *large and diverse* datasets. Specifically, CelebA-HQ is a high-quality version of the original CelebA dataset [32] which is created by applying multiple image restoration techniques on the whole CelebA dataset [68] and a subset of 30K images is selected using a frequency-based quality metric that favors images with a wider frequency spectrum to ensure large diversity. Then the FFHQ dataset [68] was introduced as an improvement to the CelebA-HQ dataset *w.r.t.* diversity. Specifically, FFHQ contains 70K high-quality face images, which includes a considerably larger variety, in terms of ethnicity, age, accessories (hats, sunglasses, etc.), and background. However, although effort was made to improve the diversity of these datasets, significant biases still exist in trained generators, as seen in Fig.2.

- **Observation 2:** The amount of bias *w.r.t.* a specific attribute, *e.g.*, Gender, varies in different GANs. For example, ProGAN is 60% biased towards Male, and StyleGAN2 is 42% biased towards Male. In addition, for the same GAN, the amount of bias *w.r.t.* different SAs is different, *e.g.*, in StyleGAN2, the amount of bias for *label0* of Moustache, and Young attributes is 77%, and 30%, respectively. This reveals that the setups in the previous studies [28], [38] have a mismatch with SOTA GAN, as they only focus on the extreme degree of biases *i.e.*, {90%, 80%} in their experiments. These variations in the amount of bias motivate us to increase the scale of our experiments and evaluate the performance of our proposed method for a more real-world applicable range of bias values (see Sec. V for more details).

Our analysis highlights the importance of addressing bias mitigation not only when training a GAN from scratch but also on existing pre-trained GANs. In what follows, we discuss transfer learning-based methods to address biases.

IV. PROPOSED METHODS

In this section, we outline the details of our proposed methods. First, in setup 1, we assume the availability of a large biased dataset $\mathcal{D}_{bias}$, and a small fair dataset $\mathcal{D}_{ref}$ for training a fair GAN [28]. Under this setup, we show that our proposed transfer learning approach named FAIRTL – training GAN on a large dataset followed by adaptation on a small fair dataset – can achieve very competitive results for training a fair GAN. Then, to further improve on the diversity in FAIRTL, we propose FAIRTL++ which introduces two additional innovations on top of FAIRTL. Intuitively, our FAIRTL/FAIRTL++ learns expressive knowledge during training on the large (biased) dataset. Then, this knowledge is transferred to the target model during adaptation, capturing the underlying fair distribution of the small dataset. The efficacy of the adaptation step in capturing fair distribution inspires

¹<https://github.com/NVlabs/stylegan2-ada-pytorch>

²https://github.com/facebookresearch/pytorch_GAN_zoo

³<https://github.com/DCGM/ffhq-features-dataset>

us to take one step further and propose an alternative, more challenging setup 2 where only a pre-trained GAN (potentially biased) is available, and the dataset used for training this model is *inaccessible*. As we will show in Sec. V, our proposed method is efficient under this challenging setup. We remark that the previous approach [28] cannot handle this problem setup as it relies on $\mathcal{D}_{bias}$.

A. FAIRTL: Fairness Adaptation via Transfer Learning

As seen in Fig. 1, we propose a simple transfer learning-based method in training a GAN for fair, diverse, and high-quality sample generation, based on $\mathcal{D}_{bias}$ and $\mathcal{D}_{ref}$. 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 simply train a GAN with adversarial loss [1]. 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 $\mathcal{D}_{bias} \cup \mathcal{D}_{ref}$. Next, in the adaptation stage, starting from the trained G_s and D_s , and using the same loss function, we adapt (fine-tune) [71] the generative model to learn fair SA distribution using $\mathcal{D}_{ref}$ only:

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

Here, G_t , D_t are generators and discriminators in the adaptation stage, trained on samples *only* in $\mathcal{D}_{ref}$, and z is random noise sampled from a Gaussian noise distribution p_z . Our experimental results show that this simple approach can be considered as a *strong baseline* for fair GAN training which achieves competitive performance against the SOTA method proposed in [28]. In addition, in stark contrast with [28] which requires the availability of $\mathcal{D}_{bias}$ for training a GAN, our proposed FAIRTL can be applied to a more challenging setup when we only have access to a pre-trained GAN while dataset used to train the GAN is inaccessible. This enables us to harness the representative power of pre-trained SOTA GANs (e.g., StyleGAN2 [5]) for attaining a fair generative model through bias mitigation (more discussion and results in Sec. V).

B. FAIRTL++: Quality-Aware Fairness Adaptation

One aspect to improve FAIRTL, is to preserve the pre-trained GAN's diversity when fine-tuning on $\mathcal{D}_{ref}$. Specifically, as $\mathcal{D}_{ref}$ is small, fine-tuning on it can result in forgetting the general knowledge learned during pre-training and degrade image quality and diversity [44], [57]. To prevent this, we propose FAIRTL++, which includes two additional modules when adapting to $\mathcal{D}_{ref}$: i) aligned feature adaptation to adapt SA-specific knowledge before adapting the entire knowledge encoded in GAN, ii) additional discriminator to continue to receive feedback from D_s during adaptation, to maintain image generation quality and diversity. In what follows we discuss these modules in more detail.

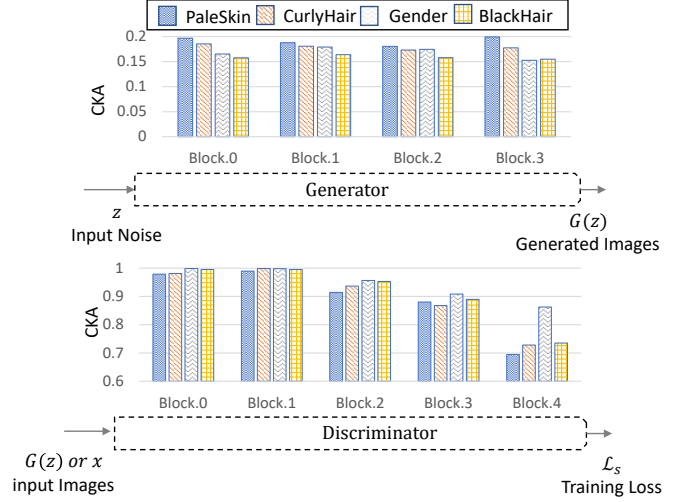


Fig. 3. **Identify the sensitive attribute-specific layers of BIGGAN, by measuring the Centered Kernel Alignment(CKA) between the same layers (residual blocks) before (i.e., pre-trained) and after adaptation.** To ensure high-quality samples after fairness adaptation, we utilize a large training dataset i.e., $perc = |\mathcal{D}_{ref}|/|\mathcal{D}_{bias}|=1.0$. Note that we use large $\mathcal{D}_{ref}$ only in this analysis. Smaller CKA values indicate more differences between the respective layers before and after adaptation, suggesting that they are SA-specific layers. Here, the layer closest to the output of the generator is block3, and that closest to the input of the discriminator is block0. Also, block0 is the layer closest to the noise input z of the generator and block4 to the output of the discriminator.

1) *Aligned Feature Adaptation:* During pre-training in FAIRTL, the GAN utilizes a large dataset to handle the task of high-quality and diverse image generation. The knowledge about this task is encoded in G_s , and D_s , without consideration for fairness. Ideally, during adaptation, we want to improve the fairness of the GAN without degrading the generally useful knowledge. More specifically, considering the SA in the generated images, the encoded knowledge in G_s , and D_s can be divided into two parts: general knowledge, and SA-specific knowledge. E.g., considering the SA Gender when generating human faces, a part of the knowledge is general regardless of Gender (e.g., eye placement is shared for Male and Female), and another part that is SA-specific (e.g., beard is specific for Male).

As knowledge is encoded in parameters of G_s and D_s , during adaptation to $\mathcal{D}_{ref}$, we aim to first modify the SA-specific parameters and then align both general and SA-specific knowledge. A similar approach is previously used for transfer learning in discriminative learning, termed Linear-Probing then Fine-Tuning (LP-FT). LPFT [55] demonstrates that when adapting a pre-trained classifier to a new task, it is advantageous to first use Linear-Probing (updating the classifier's new *task-specific* head but freezing remaining layers) for some limited epochs T , followed by Fine-Tuning (updating all layers). Experimental results in [55] suggest that LP-FT improves performance for transfer learning in discriminative learning, as it allows for more task-specific parameters to adapt, for better alignment before Fine-Tuning. Inspired by these findings, we explore applying a variant of LP-FT to generative models. Specifically, during adaptation to the fair dataset $\mathcal{D}_{ref}$, we propose to first adapt SA-specific layers of

TABLE I

Empirical evaluation of the ideal linear-probing configurations for SA adaptation. WE CONDUCT LINEAR-PROBING WITH DIFFERENT FROZEN RESIDUAL BLOCK COMBINATIONS USING BIGGAN [3] ON SA Gender. HERE, $D[m]D[n]$ INDICATES THE RESBLOCKS THAT ARE FROZEN IN D_t/G_t DURING ADAPTATION, WHERE m AND n ARE RANGES. FOR EXAMPLE, $D[0-1]G[3]$ INDICATES DISCRIMINATOR: RESBLOCK0, RESBLOCK1 AND GENERATOR: RESBLOCK3 ARE FROZEN. WE FOUND $D[0-1]G-$ TO BE THE IDEAL SETUP, CONSISTENT WITH OUR LAYER WEIGHT CHANGE EXPERIMENT IN FIG. 3.

Setup	Pre-trained	D-G-	D[0]G-	D[0-1]G-	D[0-2]G-	D[0-3]G-	D[0-1]G[3]	D[0-1]G[3-2]	D[0-1]G[3-1]
FID($\downarrow$)	25.82 $\pm$ 0.14	24.13 $\pm$ 0.12	<u>24.00</u> $\pm$ 0.10	22.90 $\pm$ 0.08	24.30 $\pm$ 0.11	24.25 $\pm$ 0.13	24.18 $\pm$ 0.14	24.36 $\pm$ 0.10	24.96 $\pm$ 0.16
FD($\downarrow$)	0.543 $\pm$ 0.128	<u>0.509</u> $\pm$ 0.041	0.512 $\pm$ 0.042	<u>0.509</u> $\pm$ 0.031	0.508 $\pm$ 0.043	0.511 $\pm$ 0.033	0.511 $\pm$ 0.034	0.517 $\pm$ 0.021	0.527 $\pm$ 0.075

the GAN for some limited epochs T and then all the layers to get a better alignment between fairness, image quality, and sample diversity. To do this, first, we need to identify the SA-specific layers.

Identifying Sensitive Attribute-Specific Layers. The gold standard for adapting FAIRTL is when $|\mathcal{D}_{ref}|$ is large enough (equal size as $|\mathcal{D}_{bias}|$ in the best case) to capture the fair distribution without distorting the general knowledge. Under this ideal condition, we expect the following behavior in the network layers: layers that encode general knowledge are previously well learned and therefore should remain relatively constant, while SA-specific layers that encode fairness should experience significant changes due to the difference in the fairness distribution. Therefore, to identify the SA-specific layers, we conduct our study under these ideal conditions – using a large $\mathcal{D}_{ref}$ – and evaluate the changes during fairness adaptation.

To identify the SA-specific layers, we use the Centered Kernel Alignment (CKA) [72] which measures the similarity between two features’ representation. As input, CKA takes $X \in \mathbb{R}^{m \times a_1}$ and $Y \in \mathbb{R}^{m \times a_2}$, which are the activation matrices (representations) of two layers with a_1 and a_2 neurons, respectively, evaluated on the same batch of m samples. Then, for the activation of each layer, we form the Gram matrices by $K = XX^T$ and $L = YY^T$ (each of which measures the similarity of a pair of the data points according to layer activation), and compute CKA as follows:

$$CKA(K, L) = \frac{HSIC(K, L)}{\sqrt{HSIC(K, K)HSIC(L, L)}} \quad (2)$$

where HSIC denotes the Hilbert-Schmidt independence criterion [73], [74]. Specifically, given the centering matrix $H = I_n - \frac{1}{n}\mathbb{1}\mathbb{1}^T$, first, the centered Gram matrices are computed using $K' = HKH$ and $L' = HLH$, and then used to compute HSIC as follows:

$$HSIC(K, L) = (\text{vec}(K') \cdot \text{vec}(L')) / (m-1)^2 \quad (3)$$

Note that vec denotes the vectorization operator. Intuitively, HSIC measures the similarity between centered Gram matrices. Considering Eqn. (2) and (3), CKA is invariant to isotropic scaling, and orthogonal transformation of representation [72]. Note that the output of CKA ranges from $[0, 1]$, where a smaller value indicates more differences, which in our case means that the layer is more SA-specific.

To apply CKA for SA-specific layer identification, we sample a batch of m random noises z and feed it to the pre-trained GAN (G_s , and D_s), and the adapted GAN by FAIRTL (G_t , and D_t). Then, we measure the CKA between the respective

blocks of pre-trained and adapted generator/discriminator, *i.e.*, a_1 becomes the set of all neurons in $Block_i$ of G_s/D_s and a_2 becomes the set of all neurons in $Block_i$ of G_t/D_t . In this analysis, the pre-trained GAN (G_s , and D_s) can be either trained from scratch on the biased dataset (the conventional problem setup used for studying GAN fairness in [28]) or any other pre-trained GAN (the newly proposed setup). In this experiment, we mostly focus on the conventional setup (Setup 1) [28] when discussing the results, but we also perform the same experiment for the new setup (publicly available pre-trained GANs) and briefly mention the results, as similar behavior is observed in both cases.

For conventional setup, as in previous works [28], we train a separate BIGGAN [3] on CelebA for each $SA \in \{\text{Gender, BlackHair, PaleSkin, CurlyHair}\}$, where for training a GAN for each SA, $\mathcal{D}_{bias}$ has 90% of bias for that SA. We remark that similar results are seen for both 75% and 60% bias values. As mentioned earlier, to resemble the ideal case of fairness adaptation, in this experiment, we utilize a $|\mathcal{D}_{ref}|$ equally as large as $|\mathcal{D}_{bias}|$, *i.e.*, $perc = |\mathcal{D}_{ref}|/|\mathcal{D}_{bias}| = 1$, to conduct both pre-training and adaptation using FAIRTL. For our new setup of adapting a pre-trained GAN, similar to the analysis in Sec. III, we utilize publicly available StyleGAN2 pre-trained on FFHQ and use a $|\mathcal{D}_{ref}| = 0.5|\mathcal{D}_{bias}|$ (due to disproportionate SA distribution resulting in the lack of samples in FFHQ) to perform the adaptation stage of FAIRTL for the same SAs in the previous setup. More training details can be found in Sec. V. Finally, we sample 3.6K noise vectors and save them. Then, for each setup, we feed these noise vectors into the G_t and G_s , and an additional 3.6K equally distributed real images into D_t and D_s . Then, as mentioned before, we measure the CKA between respective blocks in G_s / D_s , and G_t / D_t .

Fig. 3 illustrates the CKA results in identifying the SA-specific layers. Specifically, for the conventional setup, the analysis demonstrates that generally over multiple $SA \in \{\text{Gender, BlackHair, CurlyHair, PaleSkin}\}$, the first two blocks of the discriminator (closest to the model’s input) express the highest CKA > 0.98 . Therefore, the first two blocks of the D encode general knowledge (minimum SA-specific knowledge), while other blocks of D, and G encode more SA-specific knowledge. Similarly, in the proposed new setup 2, we found that only the first two blocks of StyleGAN2’s discriminator expressed a high CKA value of > 0.97 . These findings suggest that the first two blocks encode general knowledge, and the remaining blocks have more SA-specific knowledge. Therefore, to have a better feature alignment during the adaptation stage to $\mathcal{D}_{ref}$, we propose to freeze these two blocks for limited iterations and update the remaining

blocks, followed by updating all blocks for the remaining iterations.

To validate our above findings, we conduct additional analysis. Specifically, using the conventional setup with FAIRTL, we pre-train a BIGGAN for SA Gender with $perc = D_{ref}/D_{bias} = 0.1$ (train on $D_{bias} \cup D_{ref}$ as mentioned before), and then adapt the model using D_{ref} for only $T=50$ epochs. We repeat this adaptation stage with different combinations of frozen blocks for both G, and D. Then, for pre-trained G_s , and adapted G_t with only limited epochs, we evaluate the quality and diversity of the generated samples with FID [75], and fairness with the fairness discrepancy metric (FD) [28]. Note that for both metrics, a smaller value indicates better performance. See Sec. V for more details on the performance metrics. Results are tabulated in Tab. I, where $D[m]G[n]$ denotes the range of blocks frozen during adaptation, e.g., $D[0-1]G[3]$ indicates Blocks 0, and 1 of D, and Block 3 of G are frozen during adaptation. These results confirm our findings in Fig. 3 that freezing the first two layers of the discriminator ($D[0-1]G-$) leads to the best improvements in FID, while still maintaining similar improvements in FD as the baseline $D-G-$ (no frozen blocks during adaptation). Similar results were observed for the other SAs, as well as another setup. We remark that these findings also align with works from domain adaptation [44], however, to the best of our knowledge, the concept of SA-specific knowledge pertaining to fairness adaptation is proposed only in our work. Therefore, when adapting G_s and D_s to the fair dataset D_{ref} , we first freeze the lower blocks of the discriminator for some limited epochs and then fine-tune all parameters. Inspired by [55], for simplicity, we call this adaptation approach LP-FT.

2) *Balancing Fairness and Quality Objectives*: During FAIRTL, our GAN tries to achieve fair SA distribution while maintaining the high-quality samples learned during pre-training. However, we found that a single feedback signal from one discriminator may be insufficient to simultaneously convey both objectives. Other works [59], [64] had proposed the utilization of collective knowledge from multiple pre-trained discriminators to improve GAN performance, specifically 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 adopt a *multiple feedback* (MF) 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. 4 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, while making λ too big limits the feedback on the sample's quality. Empirically,

we set $\lambda = 0.6$ for all our experiments.

$$\min_{G_t} \max_{D_t} \mathcal{L} = \mathbb{E}_{x \in D_{ref}} [\log D_t(x)] + \lambda \mathbb{E}_{z \sim p_z(z)} [\log (1 - D_t(G_t(z)))] + (1 - \lambda) \mathbb{E}_{z \sim p_z(z)} [\log (1 - D_s(G_t(z)))] \quad (4)$$

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

TABLE II

Comparing our proposed FAIRTL/FAIRTL++ against Imp-weighting [28] on CelebA [32], for single SA (Gender). WE UTILIZE A D_{bias} WITH $bias = \{90_10, 75_25, 60_40\}$ WHERE 90_10 INDICATES THAT D_{bias} CONTAIN 90% FEMALE AND 10% MALE SAMPLES. THEN, WE VARIED THE SAMPLE SIZE OF D_{ref} , WHILE $|D_{bias}|$ IS KEPT CONSTANT. THIS IS DENOTED BY $perc = |D_{ref}|/|D_{bias}|$ FOR $\{0.25, 0.1, 0.05, 0.025\}$. WITH THIS SETUP, WE UTILIZE BIGGAN [3] TO REPRODUCE (A) IMP-WEIGHTING [28] 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 $perc$ AND $bias$, WHILE FAIRTL++ DEMONSTRATES EVEN GREATER IMPROVEMENTS. FOR FID AND FD, A LOW SCORE INDICATES HIGHER QUALITY SAMPLES AND FAIRER SA DISTRIBUTION, RESPECTIVELY.

90_10				
Perc	0.25	0.1	0.05	0.025
a) Imp-weighting [28]				
FID(↓)	19.20 ± 0.10	20.42 ± 0.20	23.01 ± 0.15	25.82 ± 0.13
FD (↓)	0.090 ± 0.011	0.107 ± 0.022	0.167 ± 0.016	0.246 ± 0.032
b) FAIRTL				
FID(↓)	14.21 ± 0.02	20.00 ± 0.10	22.99 ± 0.09	23.60 ± 0.11
FD (↓)	0.087 ± 0.007	0.105 ± 0.020	0.107 ± 0.012	0.130 ± 0.029
c) FAIRTL++				
FID(↓)	9.02 ± 0.03	10.69 ± 0.11	20.12 ± 0.04	20.70 ± 0.08
FD (↓)	0.010 ± 0.007	0.062 ± 0.022	0.035 ± 0.034	0.092 ± 0.025
75_25				
a) Imp-weighting [28]				
FID(↓)	24.75 ± 0.14	22.20 ± 0.17	23.94 ± 0.11	46.79 ± 0.21
FD (↓)	0.090 ± 0.010	0.322 ± 0.022	0.337 ± 0.026	0.538 ± 0.034
b) FAIRTL				
FID(↓)	19.42 ± 0.09	22.80 ± 0.10	23.42 ± 0.13	20.61 ± 0.17
FD (↓)	0.091 ± 0.011	0.149 ± 0.014	0.124 ± 0.019	0.175 ± 0.013
c) FAIRTL++				
FID(↓)	17.03 ± 0.07	22.80 ± 0.11	23.07 ± 0.07	20.12 ± 0.14
FD (↓)	0.062 ± 0.013	0.099 ± 0.021	0.117 ± 0.015	0.160 ± 0.009
60_40				
a) Imp-weighting [28]				
FID(↓)	24.85 ± 0.12	24.26 ± 0.13	23.15 ± 0.11	44.44 ± 0.15
FD (↓)	0.254 ± 0.024	0.286 ± 0.018	0.266 ± 0.019	0.487 ± 0.021
b) FAIRTL				
FID(↓)	19.76 ± 0.06	21.13 ± 0.11	22.78 ± 0.12	22.94 ± 0.14
FD (↓)	0.128 ± 0.01	0.109 ± 0.012	0.153 ± 0.014	0.149 ± 0.016
c) FAIRTL++				
FID(↓)	15.54 ± 0.09	17.88 ± 0.16	22.74 ± 0.09	21.67 ± 0.12
FD (↓)	0.108 ± 0.04	0.076 ± 0.022	0.130 ± 0.023	0.148 ± 0.011

V. EXPERIMENTS

In this section, we evaluate the performance of the proposed FAIRTL and FAIRTL++ for fair image generation. For

TABLE III

COMPARING OUR PROPOSED FAIRTL/FAIRTL++ AGAINST IMP-WEIGHTING [28] ON UTKFACE [76], FOR *single SA* (Race-Caucasian). WE UTILIZE THE SAME *single SA* SETUP, AS PER TAB. II. GIVEN THAT UTKFACE IS A SMALL DATASET, WE ARE LIMITED TO $perc = \{0.25, 0.1\}$. WE THEN SIMILARLY UTILIZE BIGGAN AND COMPARE A) [28] AS THE CURRENT SOTA AGAINST OUR PROPOSED B) FAIRTL AND C) FAIRTL++. SIMILARLY, WE FIND THAT OUR PROPOSED SOLUTIONS OUTPERFORM IMP-WEIGHTING IN BOTH FID AND FD.

Perc	90_10		75_25		60_40	
	0.25	0.1	0.25	0.1	0.25	0.1
a) Imp-weighting [28]						
FID(↓)	27.57 ± 0.45	39.03 ± 0.72	24.71 ± 0.23	39.43 ± 0.32	26.89 ± 0.45	45.34 ± 0.55
FD (↓)	0.154 ± 0.031	0.205 ± 0.044	0.128 ± 0.035	0.229 ± 0.041	0.188 ± 0.024	0.193 ± 0.022
b) FAIRTL						
FID(↓)	20.70 ± 0.32	22.92 ± 0.22	22.11 ± 0.33	24.89 ± 0.40	26.22 ± 0.36	28.76 ± 0.35
FD (↓)	<u>0.044</u> ± 0.017	<u>0.039</u> ± 0.015	<u>0.104</u> ± 0.021	<u>0.142</u> ± 0.025	<u>0.139</u> ± 0.020	<u>0.158</u> ± 0.029
c) FAIRTL++						
FID(↓)	19.21 ± 0.32	21.22 ± 0.19	21.04 ± 0.24	23.99 ± 0.22	26.83 ± 0.31	27.76 ± 0.29
FD (↓)	0.018 ± 0.020	0.003 ± 0.002	0.092 ± 0.019	0.122 ± 0.031	0.123 ± 0.033	0.149 ± 0.039

TABLE IV

COMPARING OUR PROPOSED FAIRTL/FAIRTL++ AGAINST IMP-WEIGHTING [28] ON CELEBA [32], FOR MULTI-SA

({Gender, Blackhair}). FOR *multi-SA* A $\mathcal{D}_{bias}$ IS USED WITH A BIAS OF 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 SIMILAR TO TAB. II, WE COMPARE IMP-WEIGHTING, FAIRTL AND FAIRTL++ AT DIFFERENT *perc*.

Perc	Multi			
	0.25	0.1	0.05	0.025
a) Imp-weighting [28]				
FID(↓)	14.61 ± 0.21	16.92 ± 0.31	19.43 ± 0.23	22.80 ± 0.13
FD (↓)	0.142 ± 0.032	0.116 ± 0.020	0.135 ± 0.014	0.144 ± 0.016
b) FAIRTL				
FID(↓)	<u>11.98</u> ± 0.12	<u>13.10</u> ± 0.14	<u>13.29</u> ± 0.16	<u>13.99</u> ± 0.10
FD (↓)	<u>0.113</u> ± 0.021	<u>0.115</u> ± 0.017	<u>0.118</u> ± 0.013	<u>0.138</u> ± 0.011
c) FAIRTL++				
FID(↓)	10.50 ± 0.10	11.38 ± 0.11	12.00 ± 0.10	13.18 ± 0.06
FD (↓)	0.016 ± 0.010	0.090 ± 0.020	0.086 ± 0.020	0.101 ± 0.016

this evaluation, we use two different setups: **Setup 1**) the conventional setup of [28] where we assume the availability of both a large biased dataset ($\mathcal{D}_{bias}$) and a small fair dataset ($\mathcal{D}_{ref}$), **Setup 2**) a more challenging setup introduced in this work, where only a pre-trained GAN and the small fair dataset ($\mathcal{D}_{ref}$) is available (the large dataset used to pre-train the GAN is inaccessible). In this setup, we aim to debias the already pre-trained GAN for fair image generation using $\mathcal{D}_{ref}$. Finally, we conduct an ablation study to justify our performance gains over the existing SOTA, analyze the existing effect of FAIRTL / FAIRTL++ on non-target SA and, analyze the contributions of LP-FT and multi-feedback on FAIRTL++. To demonstrate, the improved performance of our proposed approaches, for setup 1 we compare our proposed method against Imp-weighting [28] which produces SOTA performance in quality and fairness. Then, as Imp-weighting cannot be applied to the second setup due to the unavailability of the large dataset $\mathcal{D}_{bias}$, we compare the performance of the proposed method in mitigating the existing bias, with each SOTA pre-trained GAN as the baseline. 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 $\mathcal{D}_{ref}$ and $\mathcal{D}_{bias}$, when available.

Evaluation Metric. Following [28] to evaluate the quality and diversity of our generated samples we utilize FID [75], and to evaluate fairness we utilize the fairness discrepancy metric (FD) [28] defined by:

$$f = |\bar{p} - \hat{p}|_2 \quad (5)$$

Here, the FD is the L2 norm between a uniform distribution, $\bar{p} = [0.5, 0.5]$, and the SA distribution measured by the classifier, $\hat{p} = [\hat{p}_0, \hat{p}_1]$. As mentioned in Sec. III, for measuring $\hat{p}$, following [28] we use a SA classifier trained on the labeled data. A smaller FD would indicate a fairer generator *w.r.t.* the SA. For evaluating FID, similar to [28], 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 ideal generator with fair SA distribution. A smaller FID would imply better sample quality and diversity. Furthermore, we can also compare the change in FID scores before and after SA adaptation as an indicator of the generator's ability to preserve other (non-target) SA.

A. Setup 1: Training a Fair Generator

Utilizing the setup from [28], we implement our proposed method, by first training a BIGGAN [3] model with all the available data ($\mathcal{D}_{bias} \cup \mathcal{D}_{ref}$) to achieve the highest quality generator. This is followed by the adaptation stage, with $\mathcal{D}_{ref}$ only. For a fair comparison, we utilize the source code from [28] to reproduce their proposed Imp-weighting.

Dataset. We consider the datasets CelebA [32] and UTKFace [76] for this experiment. For CelebA, following Choi *et al.*, we utilize the single SA {Gender}, and multi SA {Gender, BlackHair}. Then, for UTKFace, we utilize the SA {Race(Caucasian)}. In both single attribute settings, since we cannot directly control the generator's bias,

TABLE V
EVALUATING OUR PROPOSED FAIRTL/FAIRTL++ ON A PRE-TRAINED STYLEGAN2 [5] TRAINED ON FFHQ [68] AND ProGAN [4] TRAINED ON CELEBA-HQ [77]. 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 $perc = |\mathcal{D}_{ref}|/|\mathcal{D}_{bias}|$ OF 0.025, WE IMPLEMENT FAIRTL AND FAIRTL++ AND SIMILARLY MEASURE THE DEBIASED GAN’S FID AND FD. BASED ON OUR RESULTS, WE DEMONSTRATE THAT OUR PROPOSED METHOD IS ADVANTAGEOUS ACROSS SA IN IMPROVING THE DIVERSITY, QUALITY, AND FAIRNESS OF THE GENERATED SAMPLES *w.r.t.* THE SA.

	SA	Gender	BlackHair	Young	Smiling	Moustache
StyleGAN2 (FFHQ)						
Pre-trained	FID(↓)	9.20 ± 0.02	14.58 ± 0.11	24.60 ± 0.21	9.30 ± 0.03	19.84 ± 0.21
	FD (↓)	0.102 ± 0.019	0.075 ± 0.002	0.277 ± 0.012	0.168 ± 0.007	0.376 ± 0.041
FAIRTL	FID(↓)	<u>9.01 ± 0.01</u>	<u>13.39 ± 0.09</u>	<u>12.94 ± 0.10</u>	<u>9.15 ± 0.01</u>	<u>13.03 ± 0.14</u>
	FD (↓)	<u>0.088 ± 0.010</u>	<u>0.058 ± 0.012</u>	<u>0.093 ± 0.023</u>	<u>0.098 ± 0.020</u>	<u>0.096 ± 0.042</u>
FAIRTL++	FID (↓)	8.81 ± 0.01	12.32 ± 0.10	11.79 ± 0.12	8.90 ± 0.02	11.66 ± 0.14
	FD (↓)	0.067 ± 0.014	0.057 ± 0.008	0.056 ± 0.011	0.061 ± 0.023	0.025 ± 0.028
ProGAN (CelebA-HQ)						
Pre-trained	FID(↓)	32.07 ± 0.11	32.20 ± 0.16	33.95 ± 0.09	30.41 ± 0.12	42.97 ± 0.25
	FD (↓)	0.144 ± 0.022	0.068 ± 0.012	0.053 ± 0.011	0.030 ± 0.009	0.198 ± 0.032
FAIRTL	FID(↓)	<u>29.70 ± 0.08</u>	<u>31.80 ± 0.20</u>	<u>33.86 ± 0.11</u>	<u>30.12 ± 0.12</u>	<u>39.41 ± 0.29</u>
	FD (↓)	<u>0.143 ± 0.015</u>	<u>0.024 ± 0.008</u>	<u>0.004 ± 0.003</u>	<u>0.026 ± 0.006</u>	<u>0.026 ± 0.019</u>
FAIRTL++	FID(↓)	29.49 ± 0.07	31.64 ± 0.12	30.04 ± 0.09	30.11 ± 0.09	37.99 ± 0.35
	FD (↓)	0.076 ± 0.017	0.008 ± 0.010	0.003 ± 0.03	0.015 ± 0.009	0.007 ± 0.006

we synthetically introduce a large range of data bias *i.e.*, $bias \in \{90_10, 75_25, 60_40\}$ to $\mathcal{D}_{bias}$. *E.g.*, for SA Gender, if $bias=90_10$, then $\mathcal{D}_{bias}$ contains 90% Female and 10% Male samples. For multi-attribute settings, given the data limitations, we similarly introduce a $\mathcal{D}_{bias}$ 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), Black-Hair(BH) and no-BlackHair(NBH). Next, regarding the size of $\mathcal{D}_{ref}$, we considered different $perc = |\mathcal{D}_{ref}|/|\mathcal{D}_{bias}|$, where $|\mathcal{D}_{bias}|$ is kept constant. This allows us to evaluate the robustness of our proposed method with different sample sizes, during adaptation. For the CelebA dataset, we explore $perc \in \{0.25, 0.1, 0.05, 0.025\}$, and for UTKFace, due to its smaller dataset, we explore $perc \in \{0.25, 0.1\}$.

Single Attribute Results. Tab. II presents the results of Imp-weighting [28] against our proposed methods on the CelebA dataset. Comparing across different $perc$ and $bias$, 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, *e.g.*, $perc=0.025$, FAIRTL++ can achieve a relatively fair generator, *i.e.*, low FD measurement, while Imp-weighting worsens under these conditions. This is particularly more apparent at lower $bias$ and $perc$, *e.g.*, $bias=60_40$ and, $perc=0.025$. We attribute the poor performance of Imp-weighting to the inaccuracies in the density ratio classifier, more details in Sec. V-C. Additionally, the improvement in the FID for FAIRTL and FAIRTL++, after fairness adaptation, is indicative that the diversity of the generated samples *i.e.*, (non-target) SAs representations are preserved. See Sec. V-C for a more detailed

study. Similarly, Tab. III tabulates results on the UTKFace dataset with SA Race-Caucasian. With this dataset, we similarly observe that FAIRTL++ outperforms both Imp-weighting and FAIRTL in quality, diversity and fairness. In fact, on the smaller UTKFace dataset, the benefits of our proposed method become more prominent with Imp-weighting’s sample quality /diversity (FID) significantly degrading under smaller $perc$ and $bias$. In contrast, our proposed method only experiences minor degradation, while still enforcing SOTA fairness (lower FD).

Multi-Attribute Results. Tab. IV presents a similar experiment but with considering multiple SA: Gender and BlackHair. Our results show that even under this more challenging setup, involving two SA simultaneously, FAIRTL++ still outperforms Imp-weighting, achieving SOTA performance in fairness and sample quality.

B. 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 ($\mathcal{D}_{bias}$). Instead, we can improve the fairness of existing biased pre-trained models, by just performing the adaptation stage of FAIRTL / FAIRTL++ on $\mathcal{D}_{ref}$. One major advantage of this setup is the significant reduction in required computational resources to attain a fair generative model. For this experiment, as per Sec III, we utilize the publicly available pre-trained StyleGAN2 [5] on the FFHQ dataset [68] and ProGAN [4] on CelebA-HQ [77]. As mentioned earlier, our work is the first to propose this setup, and therefore there are no existing baselines to compare against. More specifically, the existing SOTA —Imp-weighting [28]— requires training from scratch, more details in Sec. V-C. Therefore, we adopt the next most appropriate baseline and evaluate FAIRTL /

FAIRTL++ performance in debiasing the already existing pre-trained (potentially biased) SOTA GANs. To do this, we first measure the FID and FD of the pre-trained model across different SAs. Then using $\mathcal{D}_{ref}$ we implement the adaptation stage for FAIRTL / FAIRTL++, followed by re-evaluating the models to determine their performance.

Dataset. As discussed in Sec. III, we utilize the FFHQ and CelebA-HQ dataset for StyleGAN2 and ProGAN respectively, and consider the $SA \in \{\text{Gender, BlackHair, Young, Smiling, Moustache}\}$ to demonstrate the effectiveness of our proposed method across different SAs. For each SA, we attained a $\mathcal{D}_{ref}$ with $perc = 0.025$.

From our results in Tab. V, as discussed earlier, we observe that both the pre-trained StyleGAN2 and ProGAN models contain a considerable amount of bias in the selected SA. In particular, larger biases exist for $\{\text{Young, Smiling, Moustache}\}$ in StyleGAN2, and for $\{\text{Gender, Moustache}\}$ in ProGAN, where high FD measurements were observed. Furthermore, the higher FID measurements indicate a mismatch between the quality/diversity of the generated samples and the ideal reference samples. Our proposed solutions however proves to be effective in improving both the fairness and quality of both GANs as seen from the lower FD and FID score. Similar to the previous experiments, FAIRTL++ proves to be the more effective method. Fig. 4 provides some qualitative analysis, where we illustrate a few random samples that have been adapted from the majority-represented class of some SAs to the minority-represented class in the respective SA in StyleGAN2, thereby achieving a fairer SA distribution. We remark that although the SA of the samples has been adapted *e.g.*, Female $\rightarrow$ Male, the underlying general (non-target) SA *e.g.*, pose remains unchanged, thereby preserving diversity as indicated by the FID measurements.

C. Analysis and Ablation Study

Justification of the performance gain with our approach.

As we will discuss below, our approach achieves improved quality and fairness as it departs from using the *density ratio classifier* (DRC), which was central to Imp-weighting [28] but contains some major drawbacks. Specifically, the DRC-based method [28] takes both $\mathcal{D}_{bias}$ and $\mathcal{D}_{ref}$ samples as inputs, and learns to classify the samples to their respective datasets. By doing so, the DRC implicitly learns to differentiate between the minority and majority labels of the SA. Then during GAN training, the trained DRC’s softmax output is utilized to re-weight the loss function for samples belonging to $\mathcal{D}_{bias}$:

$$\mathcal{L}(\theta, \mathcal{D}_{bias}, \mathcal{D}_{ref}) = \mathbb{E}_{x_i \in \mathcal{D}_{ref} \cup \mathcal{D}_{bias}} w(x_i) l(x_i, \theta),$$

$$\text{where } w(x_i) = \begin{cases} 1 & x_i \in \mathcal{D}_{ref} \\ \frac{p_{ref}(x_i)}{p_{bias}(x_i)} & x_i \in \mathcal{D}_{bias} \end{cases} \quad (6)$$

Here, $p_{ref}(x_i)/p_{bias}(x_i)$ is the ratio of the density ratio classifier’s probabilistic output that a sample x_i belongs to $\mathcal{D}_{ref}$, to the probability that it belongs to $\mathcal{D}_{bias}$. One concern for this algorithm is that the DRC is not always accurate, particularly when $|\mathcal{D}_{ref}|$ is small. This results in poor estimation of $w(x_i)$. To analyze this, we train the DRC following the procedures of [28]. From our results in Fig.5, we notice that as the

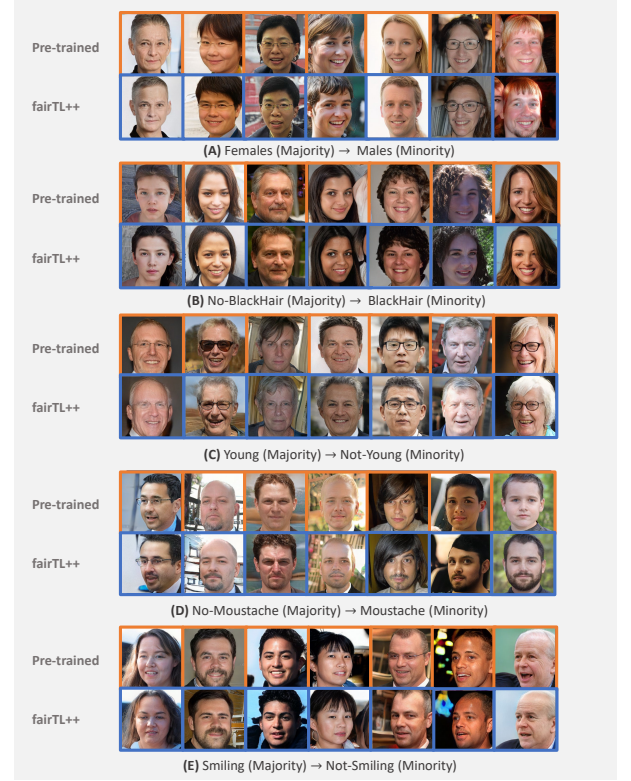


Fig. 4. **Illustration of samples before and after fairness adaptation by our FAIRTL++ on a pre-trained StyleGAN2 [5].** For each sample, we utilize the same noise vector to sample from the pre-trained model and FAIRTL++ after SA adaptation. Notice how our methods can enable sample adaptation from majority to minority-represented SA, while general features *e.g.*, pose are preserved.

$|\mathcal{D}_{ref}|$ decreases, the calibration of the DRC worsens *i.e.*, deviates from the ideal diagonal calibration curve. As a result, the DRC provides inaccurate re-weighting, resulting in the worsening of the sample quality and fairness, seen in Sec. V-A. On the other hand, our approach takes a transfer learning approach, departing from DRC-based design and eliminating the inaccuracy issue of DRC. This leads to improved quality and fairness as demonstrated in our experiment results.

Analyzing the effect of fairness adaptation on (non-target) SA. We provide a more explicit evaluation to determine if implementing fairness adaptation on a selected SA would impact other (non-target) SA. To do this, we utilize the same setup in Tab. V and assess the FD of the remaining (non-target) SA $\{\text{Gender, BlackHair, Young, Smiling, Moustache}\}$ before and after implementing FAIRTL++ on the target SA.

Our results in Tab. VI, demonstrate that generally, all the other (non-target) SA biases remain relatively constant after fairness adaptation. However, there exist some SA that may be more entangled with others *e.g.*, Gender and Moustache which observe some perturbation to their FD when the other is adapted as the target SA.

Analyzing the effects of Linear Probing then Fine-tuning (LP-FT) and Multi-feedback (MF). Utilizing the same setups as Sec. V-A we evaluate the individual improvements brought by the addition of LP-FT and MF to FAIRTL++. To do this, we utilize FAIRTL with either LP-FT or MF.

TABLE VI

ABLATION STUDY TO EVALUATE THE BIAS OF (NON-TARGETED) SENSITIVE ATTRIBUTES (SAS), BEFORE AND AFTER IMPLEMENTING FAIRTL++ *w.r.t.* A TARGET SA. FOR THIS EXPERIMENT, WE USE SETUP2 AND ADAPT A PRE-TRAINED STYLEGAN2 [5] TRAINED ON FFHQ TO A TARGET SA $\in \{\text{Gender, BlackHair, Young, Smiling, Moustache}\}$, AND MEASURE FAIRNESS *w.r.t.* OTHER SAS USING FD METRIC. FOR FAIRTL++, WE SIMILARLY UTILISE A $\text{perc} = |\mathcal{D}_{ref}|/|\mathcal{D}_{bias}|$ OF 0.025. THE FD VALUES OF THE PRE-TRAINED GENERATOR *w.r.t.* DIFFERENT SAS ARE SHOWN IN THE FIRST ROW OF THE NUMBERS. FOR THE REMAINING ROWS UNDER FAIRTL++, EACH ROW SHOWS THE FD VALUES AFTER ADAPTING THE GENERATOR TO A TARGET SA. FOR EXAMPLE, THE SECOND ROW OF NUMBERS IN THE TABLE SHOWS THE RESULTS OF ADAPTING THE GENERATOR TO Gender USING FAIRTL++. IN THIS ROW, AS ONE CAN SEE THE FD FOR THE Gender DECREASES ($0.102 \pm 0.019 \rightarrow 0.067 \pm 0.014$), WHILE AN UNRELATED SA LIKE Young REMAIN RELATIVELY CONSTANT ($0.277 \pm 0.012 \rightarrow 0.280 \pm 0.009$).

Targeted SA		FD ($\downarrow$)				
		Gender	BlackHair	Young	Smiling	Moustache
Pre-trained	-	0.102 ± 0.019	0.075 ± 0.002	0.277 ± 0.012	0.168 ± 0.007	0.376 ± 0.041
FAIRTL++	Gender	0.067 ± 0.014	0.072 ± 0.005	0.280 ± 0.009	0.165 ± 0.010	0.342 ± 0.036
	BlackHair	0.104 ± 0.021	0.057 ± 0.008	0.270 ± 0.019	0.174 ± 0.012	0.371 ± 0.049
	Young	0.108 ± 0.011	0.078 ± 0.014	0.056 ± 0.011	0.168 ± 0.010	0.380 ± 0.025
	Smiling	0.110 ± 0.016	0.077 ± 0.007	0.278 ± 0.010	0.061 ± 0.023	0.370 ± 0.044
	Moustache	0.051 ± 0.022	0.079 ± 0.004	0.281 ± 0.019	0.164 ± 0.011	0.025 ± 0.028

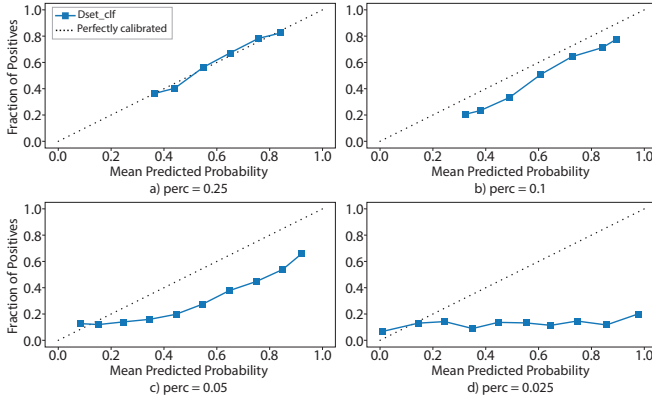


Fig. 5. Analyzing accuracy of Density Ratio Classifiers (DRCs) as adopted in previous DRC-based approach [28], at different perc . We trained the DRC, a ResNet-18 [78] classifier, to differentiate between $\mathcal{D}_{ref}$ and $\mathcal{D}_{bias}$, with 90%-10% bias on SA Gender. We utilized different $\text{perc} = \{0.25, 0.1, 0.05, 0.025\}$ and plot the calibration curve: mean predicted probability (x-axis) vs the fraction of the positive prediction of a batch of samples (y-axis). The diagonal trend-line depicts the ideal DRC where the classifier is perfectly calibrated.

Our results in Tab. VII for $\text{bias}=90_{-10}$, shows that both FAIRTL w/ LP-FT and FAIRTL w/ MF, demonstrate improvement in FAIRTL’s quality and fairness, as seen by the smaller FID and FD scores. Then, when combining both methods (FAIRTL++) we observe even greater improvements in FID and FD, surpassing the individual improvements made by the addition of each method. This demonstrates the compatibility of LP-FT and MF in improving the fair transfer learning process. We remark that similar observations were seen for the other setups.

VI. LIMITATIONS

Our proposed FAIRTL / FAIRTL++ helps to improve the fairness of a generative model *w.r.t.* a target SA. However, users must be careful in considering the appropriate target SA in their respective applications. In some cases, certain (non-targeted) SA may have entanglement with the target SA,

TABLE VII

ABLATION STUDY OF FAIRTL++ ON CELEBA [32], FOR SINGLE SA (Gender). WE UTILIZE A $\mathcal{D}_{bias}$ WITH $\text{bias} = 90_{-10}$ i.e., 90% SAMPLE ARE FEMALE AND 10% MALE. THEN, WE VARIED $\text{perc} = |\mathcal{D}_{ref}|/|\mathcal{D}_{bias}|$ FOR $\{0.25, 0.1, 0.05, 0.025\}$. WITH THIS SETUP, WE UTILIZE BIGGAN [3] TO IMPLEMENT (a) FAIRTL (b) FAIRTL WITH LINEAR PROBING AND FINE TUNING (LP-FT) (c) FAIRTL WITH MULTIPLE-FEEDBACK(MF) AND (d) FAIRTL++.

Perc	0.25	0.1	0.05	0.025
a) FAIRTL				
FID($\downarrow$)	14.21 ± 0.02	20.00 ± 0.10	22.99 ± 0.09	23.60 ± 0.11
FD ($\downarrow$)	0.087 ± 0.007	0.105 ± 0.020	0.107 ± 0.012	0.130 ± 0.029
b) FAIRTL w/ LP-FT				
FID($\downarrow$)	10.98 ± 0.06	19.30 ± 0.11	21.03 ± 0.07	21.6 ± 0.10
FD ($\downarrow$)	0.062 ± 0.012	0.070 ± 0.018	0.089 ± 0.021	0.126 ± 0.023
c) FAIRTL w/ MF				
FID($\downarrow$)	10.20 ± 0.07	10.85 ± 0.09	20.20 ± 0.06	21.00 ± 0.09
FD ($\downarrow$)	0.016 ± 0.013	0.062 ± 0.20	0.120 ± 0.011	0.095 ± 0.020
d) FAIRTL++				
FID($\downarrow$)	9.02 ± 0.03	10.69 ± 0.11	20.12 ± 0.04	20.70 ± 0.08
FD ($\downarrow$)	0.010 ± 0.007	0.062 ± 0.022	0.035 ± 0.034	0.092 ± 0.025

resulting in some unfavorable perturbation in the former’s representation, during fairness adaptation. We further remark that, as our approach utilizes the concept of group fairness [28], there are no theoretical limitations on the number of selected SA or the number of classes per SA, so long as an appropriate $\mathcal{D}_{ref}$ can be acquired.

VII. CONCLUSION

In our work, we focus on the important 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 our aligned feature adaptation

by identifying attribute-specific parameters, and our multiple feedback for receiving quality assessment from pre-trained discriminators during adaptation can help further improve both sample quality and fairness, thereby achieving SOTA performance. Additionally, we demonstrate that our proposed methods can similarly improve the quality and fairness of SOTA pre-trained GANs.

VIII. ACKNOWLEDGEMENTS

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programmes (AISG Award No.: AISG2-TC-2022-007); SUTD project PIE-SGP-AI-2018-01; the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021). This material is based on the research/work support in part by the Changi General Hospital and Singapore University of Technology and Design, under the HealthTech Innovation Fund (HTIF Award No. CGH-SUTD-2021-004).

REFERENCES

- [1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.
- [2] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville, "Improved Training of Wasserstein GANs," *arXiv:1704.00028 [cs, stat]*, Dec. 2017.
- [3] A. Brock, J. Donahue, and K. Simonyan, "Large Scale GAN Training for High Fidelity Natural Image Synthesis," *arXiv:1809.11096 [cs, stat]*, Feb. 2019.
- [4] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.
- [5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and Improving the Image Quality of StyleGAN," *arXiv:1912.04958 [cs, eess, stat]*, Mar. 2020.
- [6] U. Ojha, Y. Li, J. Lu, A. A. Efros, Y. Jae Lee, E. Shechtman, and R. Zhang, "Few-shot Image Generation via Cross-domain Correspondence," in *IEEE/CVF conference on computer vision and pattern recognition*, Jun. 2021, pp. 10 738–10 747.
- [7] Y. Yang, J. Zhou, J. Ai, Y. Bin, A. Hanjalic, H. T. Shen, and Y. Ji, "Video captioning by adversarial lstm," *IEEE Transactions on Image Processing*, vol. 27, no. 11, pp. 5600–5611, 2018.
- [8] K. Prajapati, V. Chudasama, H. Patel, K. Upla, K. Raja, R. Ramachandra, and C. Busch, "Direct unsupervised super-resolution using generative adversarial network (dus-gan) for real-world data," *IEEE Transactions on Image Processing*, vol. 30, pp. 8251–8264, 2021.
- [9] A. Marafioti, P. Majdak, N. Holighaus, and N. Perraudin, "Gacela: A generative adversarial context encoder for long audio inpainting of music," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 120–131, 2021.
- [10] S. U. Dar, M. Yurt, M. Shahdloo, M. E. Ildiz, B. Tinaz, and T. Çukur, "Prior-guided image reconstruction for accelerated multi-contrast mri via generative adversarial networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1072–1087, Oct 2020.
- [11] K. Mei and V. M. Patel, "Ltt-gan: Looking through turbulence by inverting gans," *IEEE Journal of Selected Topics in Signal Processing*, 2023.
- [12] H. Wang, G. Tao, J. Ma, S. Jia, L. Chi, H. Yang, Z. Zhao, and J. Tao, "Predicting the epidemics trend of covid-19 using epidemiological-based generative adversarial networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 2, pp. 276–288, Feb 2022.
- [13] J. C. Neves, R. Tolosana, R. Vera-Rodriguez, V. Lopes, H. Proença, and J. Fierrez, "Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 1038–1048, Aug 2020.
- [14] Z. Mi, X. Jiang, T. Sun, and K. Xu, "Gan-generated image detection with self-attention mechanism against gan generator defect," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 969–981, Aug 2020.
- [15] Y. Duan, C. Han, X. Tao, B. Geng, Y. Du, and J. Lu, "Panoramic image generation: from 2-d sketch to spherical image," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 194–208, 2020.
- [16] X. Chen, C. Xu, X. Yang, L. Song, and D. Tao, "Gated-gan: Adversarial gated networks for multi-collection style transfer," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 546–560, 2018.
- [17] Y. Li, S. Tang, R. Zhang, Y. Zhang, J. Li, and S. Yan, "Asymmetric gan for unpaired image-to-image translation," *IEEE Transactions on Image Processing*, vol. 28, no. 12, pp. 5881–5896, 2019.
- [18] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *IEEE/CVF conference on computer vision and pattern recognition*, 2017.
- [19] H. Tan, X. Liu, M. Liu, B. Yin, and X. Li, "Kt-gan: Knowledge-transfer generative adversarial network for text-to-image synthesis," *IEEE Transactions on Image Processing*, vol. 30, pp. 1275–1290, 2020.
- [20] H. Xu, J. Ma, and X.-P. Zhang, "Mef-gan: Multi-exposure image fusion via generative adversarial networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 7203–7216, 2020.
- [21] Y. Niu, J. Wu, W. Liu, W. Guo, and R. W. Lau, "Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions," *IEEE Transactions on Image Processing*, vol. 30, pp. 3885–3896, 2021.
- [22] Y. Li, X. Peng, J. Zhang, Z. Li, and M. Wen, "Dct-gan: Dilated convolutional transformer-based gan for time series anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [23] H. J. Jalan, G. Maurya, C. Corda, S. Dsouza, and D. Panchal, "Suspect Face Generation," in *2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA)*, Apr. 2020, pp. 73–78.
- [24] M. Frid-Adar, I. Diamant, E. Klang, M. Amitai, J. Goldberger, and H. Greenspan, "Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification," *Neurocomputing*, vol. 321, pp. 321–331, 2018.
- [25] T. Jang, F. Zheng, and X. Wang, "Constructing a Fair Classifier with Generated Fair Data," *AAAI Conference on Artificial Intelligence*, vol. 35, no. 9, pp. 7908–7916, May 2021.
- [26] A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone, and E. Ferlante, "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis," *Proceedings of the National Academy of Sciences*, vol. 117, no. 23, pp. 12 592–12 594, Jun. 2020.
- [27] E. Frankel and E. Vendrow, "Fair Generation Through Prior Modification," *Advances in Neural Information Processing Systems*, Jun. 2020.
- [28] K. Choi, A. Grover, T. Singh, R. Shu, and S. Ermon, "Fair Generative Modeling via Weak Supervision," in *International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 1887–1898.
- [29] A. I. Humayun, R. Balestriero, and R. Baraniuk, "MaGNET: Uniform Sampling from Deep Generative Network Manifolds Without Retraining," in *International Conference on Learning Representations*, Jan. 2022.
- [30] S. Tan, Y. Shen, and B. Zhou, "Improving the Fairness of Deep Generative Models without Retraining," *arXiv:2012.04842 [cs]*, Dec. 2020.
- [31] C. T. H. Teo, M. Abdollahzadeh, and N.-M. Cheung, "On Measuring Fairness in Generative Models," no. arXiv:2310.19297, Oct. 2023.
- [32] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *International Conference on Computer Vision*, Dec 2015.
- [33] V. Chizzola, B. Micheli, and Vingelli, "TARGET – Taking a Reflexive approach to Gender Equality for institutional Transformation," *STEM Gender Equality Congress Proceedings*, vol. 1, no. 1, pp. 839–839, Jun. 2017.
- [34] C. Muehlethaler and R. Albert, "Collecting data on textiles from the internet using web crawling and web scraping tools," *Forensic Science International*, vol. 322, p. 110753, 2021.
- [35] T. Le Quy, A. Roy, V. Iosifidis, W. Zhang, and E. Ntoutsi, "A survey on datasets for fairness-aware machine learning," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, p. e1452, 2022.
- [36] S. Hwang, "FairFaceGAN: Fairness-aware Facial Image-to-Image Translation," *BMVS20*, p. 14, 2020.
- [37] N. Yu, K. Li, P. Zhou, J. Malik, L. Davis, and M. Fritz, "Inclusive gan: Improving data and minority coverage in generative models," in *European Conference on Computer Vision*. Springer, 2020.
- [38] C. T. Teo, M. Abdollahzadeh, and N.-M. Cheung, "Fair generative models via transfer learning," in *AAAI Conference on Artificial Intelligence*, vol. 37, no. 2, 2023, pp. 2429–2437.
- [39] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

- [40] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 259–268.
- [41] D. Xu, S. Yuan, L. Zhang, and X. Wu, "Fairgan: Fairness-aware generative adversarial networks," in *International Conference on Big Data*. IEEE, 2018, pp. 570–575.
- [42] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness gan: Generating datasets with fairness properties using a generative adversarial network," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 3–1, 2019.
- [43] Y. ZHAO, K. Chandrasegaran, M. Abdollahzadeh, and N. man Cheung, "Few-shot image generation via adaptation-aware kernel modulation," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022.
- [44] S. Mo, M. Cho, and J. Shin, "Freeze the Discriminator: A Simple Baseline for Fine-Tuning GANs," *arXiv:2002.10964 [cs, stat]*, Feb. 2020.
- [45] Y. Zhao, C. Du, M. Abdollahzadeh, T. Pang, M. Lin, S. Yan, and N.-M. Cheung, "Exploring incompatible knowledge transfer in few-shot image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7380–7391.
- [46] M. Abdollahzadeh, T. Malekzadeh, C. T. Teo, K. Chandrasegaran, G. Liu, and N.-M. Cheung, "A survey on generative modeling with limited data, few shots, and zero shot," *arXiv preprint arXiv:2307.14397*, 2023.
- [47] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.
- [48] J. Jiang, Y. Shu, J. Wang, and M. Long, "Transferability in deep learning: A survey," *arXiv preprint arXiv:2201.05867*, 2022.
- [49] S. Wu, H. R. Zhang, and C. Ré, "Understanding and improving information transfer in multi-task learning," *arXiv preprint arXiv:2005.00944*, 2020.
- [50] T. Malekzadeh, M. Abdollahzadeh, H. Nejati, and N.-M. Cheung, "Aircraft fuselage defect detection using deep neural networks," *arXiv preprint arXiv:1712.09213*, 2017.
- [51] S. S. Du, W. Hu, S. M. Kakade, J. D. Lee, and Q. Lei, "Few-shot learning via learning the representation, provably," in *International Conference on Learning Representations*, 2020.
- [52] H. Cai, T. Wang, Z. Wu, K. Wang, J. Lin, and S. Han, "On-device image classification with proxyless neural architecture search and quantization-aware fine-tuning," in *International Conference on Computer Vision Workshops*, 2019.
- [53] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, "SpotTune: Transfer learning through adaptive fine-tuning," in *IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4805–4814.
- [54] M. Abdollahzadeh, T. Malekzadeh, and N.-M. M. Cheung, "Revisit multimodal meta-learning through the lens of multi-task learning," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [55] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution," *International Conference on Learning Representations*, Feb. 2022.
- [56] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, and B. Raducanu, "Transferring GANs: Generating Images from Limited Data," in *European Conference on Computer Vision*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, vol. 11210, pp. 220–236.
- [57] Y. Li, R. Zhang, J. C. Lu, and E. Shechtman, "Few-shot Image Generation with Elastic Weight Consolidation," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 15 885–15 896.
- [58] Y. Zhao, H. Ding, H. Huang, and N.-M. Cheung, "A closer look at few-shot image generation," in *IEEE/CVF conference on computer vision and pattern recognition*, June 2022, pp. 9140–9150.
- [59] N. Kumari, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Ensembling Off-the-Shelf Models for GAN Training," *IEEE/CVF conference on Computer Vision and Pattern Recognition*, p. 12, 2022.
- [60] N.-T. Tran, V.-H. Tran, N.-B. Nguyen, T.-K. Nguyen, and N.-M. Cheung, "On Data Augmentation for GAN Training," *IEEE Transactions on Image Processing*, vol. 30, pp. 1882–1897, 2021.
- [61] T. Nguyen, T. Le, H. Vu, and D. Phung, "Dual discriminator generative adversarial nets," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 2667–2677.
- [62] I. Durugkar, I. Gemp, and S. Mahadevan, "Generative Multi-Adversarial Networks," *International Conference on Learning Representations*, p. 14, 2017.
- [63] I. Albuquerque, J. Monteiro, T. Doan, B. Considine, T. Falk, and I. Mitliagkas, "Multi-objective training of Generative Adversarial Networks with multiple discriminators," *International Conference on Learning Representations*, Feb. 2022.
- [64] S. Um and C. Suh, "A Fair Generative Model Using Total Variation Distance," Sep. 2021.
- [65] Q. Hoang, T. D. Nguyen, T. Le, and D. Phung, "MGAN: Training Generative Adversarial Nets with Multiple Generators," in *International Conference on Learning Representations*, Feb. 2022.
- [66] A. Ghosh, V. Kulharia, V. Nambodiri, P. H. S. Torr, and P. K. Dokania, "Multi-Agent Diverse Generative Adversarial Networks," Jul. 2018.
- [67] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive Growing of GANs for Improved Quality, Stability, and Variation," *International Conference on Learning Representations*, Feb. 2018.
- [68] T. Karras, S. Laine, and T. Aila, "A Style-Based Generator Architecture for Generative Adversarial Networks," in *IEEE/CVF conference on computer vision and pattern recognition*, Jun. 2019, pp. 4401–4410.
- [69] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems*, 2017, pp. 6626–6637.
- [70] N. Liu, S. Li, Y. Du, A. Torralba, and J. B. Tenenbaum, "Compositional visual generation with composable diffusion models," in *European Conference on Computer Vision*. Springer, 2022, pp. 423–439.
- [71] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, and B. Raducanu, "Transferring gans: generating images from limited data," in *ECCV*, 2018.
- [72] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?" *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 116–12 128, 2021.
- [73] A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola, "A kernel statistical test of independence," *Advances in neural information processing systems*, vol. 20, 2007.
- [74] S. Kornblith, M. Norouzi, H. Lee, and G. Hinton, "Similarity of neural network representations revisited," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3519–3529.
- [75] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *arXiv:1706.08500 [cs, stat]*, Jan. 2018.
- [76] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," in *IEEE/CVF conference on computer vision and pattern recognition*, 2017, pp. 5810–5818.
- [77] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *IEEE/CVF conference on computer vision and pattern recognition*, 2020.
- [78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE/CVF conference on computer vision and pattern recognition*, Jun. 2016, pp. 770–778.



Christopher Teo received his BsC in Engineering and Product Development in 2018 from the Singapore Institute of Technology, and is currently a fourth-year Ph.D. student, whose research focuses on Generative Models.



Milad Abdollahzadeh received his Ph.D. in Computer and Electrical Engineering from the University of Tabriz, Tabriz, Iran in 2018. He is currently a Postdoctoral Research Fellow with the Singapore University of Technology and Design (SUTD). His research interests include machine learning and computer vision.



Ngai-Man Cheung (Senior Member, IEEE) received the Ph.D. in electrical engineering from the University of Southern California, Los Angeles, CA, USA. He is currently an Associate Professor at the Singapore University of Technology and Design. His research interests include image and signal processing, computer vision and AI.