A Fair Generative Model Using LeCam Divergence

Soobin Um[1], **Changho Suh**[2]

[1] Graduate School of AI, KAIST
[2] School of Electrical Engineering, KAIST
{sum, chsuh}@kaist.ac.kr

Abstract

We explore a fairness-related challenge that arises in generative models. The challenge is that biased training data with imbalanced demographics may yield a high asymmetry in size of generated samples across distinct groups. We focus on practically-relevant scenarios wherein demographic labels are not available and therefore the design of a fair generative model is non-straightforward. In this paper, we propose an optimization framework that regulates the unfairness under such practical settings via one statistical measure, LeCam (LC)-divergence. Specifically to quantify the degree of unfairness, we employ a balanced-yet-small reference dataset and then measure its distance with generated samples using the LC-divergence, which is shown to be particularly instrumental to a *small* size of the reference dataset. We take a variational optimization approach to implement the LC-based measure. Experiments on benchmark real datasets demonstrate that the proposed framework can significantly improve the fairness performance while maintaining realistic sample quality for a wide range of the reference set size all the way down to 1% relative to training set.

Introduction

High-quality realistic samples synthesized thanks to recent advances in generative models (Brock, Donahue, and Simonyan 2019; Goodfellow et al. 2014; Karras, Laine, and Aila 2019) have played a crucial role to enrich training data for a widening array of applications such as face recognition, natural language processing, and medical imaging (Wang, Wang, and Lian 2019; Chang, Chuang, and Lee 2018; Yi, Walia, and Babyn 2019). One challenge concerning *fairness* arises when generative models are built upon biased training data that preserve unbalanced representations of demographic groups. Any existing bias in the dataset can readily be propagated to the learned model, thus producing biased generations towards certain demographics. The unbalanced generated samples may often yield undesirable performances against underrepresented groups for downstream applications. One natural way to ensure fair sample generation is to exploit demographic labels (if available) to build a fair generative model, e.g., via conditional GANs (Mirza and Osindero 2014; Odena, Olah, and Shlens 2017; Miyato and Koyama 2018) which employ such labels to easily generate an arbitrary number of samples for minority groups. In many practically-relevant scenarios, however, such labels are often unavailable.

To address the challenge, one pioneering work (Choi et al. 2020) develops a novel debiasing technique that employs the *reweighting* idea (Ren et al. 2018; Kamiran and Calders 2012; Byrd and Lipton 2019) to put more weights to underrepresented samples, thereby promoting fair sample generation across demographic groups. One key feature of the technique is to identify the bias (reflected in the weights) via a small and unlabelled reference dataset. While it enjoys significant fairness performance for moderate sizes of the reference dataset, it may provide a marginal gain for a more practically-relevant case of a small set size where the weight estimation is often inaccurate, as hinted by the meta-learning literature (Ren et al. 2018; Shu et al. 2020). We also find such phenomenon in our experiments; see Table 1 for details.

Contribution: In this work, we take a distinct approach to address the issue w.r.t. the small reference set size. We still rely upon a balanced unlabelled reference dataset, yet employing a statistical notion, *LeCam (LC)-divergence* (Le Cam 2012), instead of the reweighting approach. One important feature of the LC-divergence was emphasized by Tseng et al. (2021) in the context of GANs. The divergence captures well the distance between real and generated samples even in the limited size of training data, thereby serving as a regularized loss in the design of a discriminator. This robustness aspect of the divergence motivates us to incorporate the LC-divergence in quantifying the degree of unfairness particularly when an employed reference set size is small. Specifically we compute the LC-divergence between reference and generated samples. We then promote fair sample generation by adding the LC-divergence as a regularization term into conventional optimization (e.g., GAN-based optimization (Goodfellow et al. 2014; Nowozin, Cseke, and Tomioka 2016; Arjovsky, Chintala, and Bottou 2017)). We employ the variational optimization technique w.r.t. the LC-divergence (Tseng et al. 2021) to translate the regularized optimization into an implementable form. We also conduct extensive experiments on three benchmark real datasets: CelebA (Liu et al. 2015), UTK-Face (Zhang, Song, and Qi 2017), and FairFace (Karkkainen

and Joo 2021). We demonstrate via simulation that the proposed framework can significantly boost up the fairness performance while offering high-quality realistic samples reflected in low Fréchet Inception Distance (Heusel et al. 2017). We also find that our approach outperforms the state of the art (Choi et al. 2020), particularly when the balanced reference set size is small: the significant improvements preserve for a wide range of the reference set size down to 1% relative to training data.

Related works: After firstly explored in Choi et al. (2020), fairness of representations in a generative model has been investigated under a number of different scenarios (Tan, Shen, and Zhou 2020; Yu et al. 2020; Jalal et al. 2021; Lee et al. 2021). For instance, Tan, Shen, and Zhou (2020) propose a different way that promotes fair sample generation by smartly perturbing the input distribution of a pre-trained generative model with the help of a classifier for sensitive attributes. The key distinction w.r.t. ours is that it relies upon the additional classifier. Another notable work that bears an intimate connection to our setting is due to Yu et al. (2020). The authors in Yu et al. (2020) employ demographic labels for minority groups to generate a wide variety of samples with improved data coverage by harmonizing GAN and MLE ideas. A distinction w.r.t. ours is that it requires the knowledge on demographic labels.

Another line of fair generative modeling focuses on *label* bias, instead of representation bias (Xu et al. 2018, 2019a,b; Sattigeri et al. 2019; Jang, Zheng, and Wang 2021; Kyono et al. 2021). The goal therein is to develop a generative model such that the generated decision labels are statistically independent of the given demographic labels. Again, these are not directly comparable to ours, as they require the use of demographic labels.

The variational optimization technique w.r.t. the LC-divergence that gives an inspiration to our work has originated from Tseng et al. (2021). The authors in Tseng et al. (2021) showed that a properly weighted LC-divergence can well represent the distance between real and generated samples, thereby serving to improve the generalization performance in the scarce training data. This finding forms the basis of our proposed framework that incorporates the weighted LC-divergence to promote the fairness reflected in a small divergence. On a different note, the LC-divergence has also been instrumental in bounding some important quantities that arise in diverse contexts such as communication complexity in theoretical computer science (Yehudayoff 2020) and growth rate in group theory (Ozawa 2015).

Problem Formulation

Setup: Figure 1 illustrates the problem setting for a fair generative model that we focus on herein. We consider a challenging yet practically-relevant scenario wherein demographic information (or that we call sensitive attribute), say $z \in \mathcal{Z}$, is not available. Under this blind setting, the goal is to construct a *fair* generative model so as to ensure the produced samples with the same size (as much as possible) across distinct demographics. We assume that there are two types of data given in the problem: (i) training data $\mathcal{D}_{\mathsf{bias}} := \{x_{\mathsf{bias}}^{(i)}\}_{i=1}^{m_{\mathsf{bias}}}$; (ii) reference data $\mathcal{D}_{\mathsf{ref}} := \{x_{\mathsf{ref}}^{(i)}\}_{i=1}^{m_{\mathsf{ref}}}$. Since training data is potentially biased, we use the word "bias" in the associated notations. Here m_{bias} denotes the number of training examples. Let $\mathbb{P}_{\mathsf{bias}}$ be the probability distribution which each training data $x_{\mathsf{bias}}^{(i)} \in \mathcal{X}$ is drawn from. In a biased scenario having female-vs-male sensitive attribute, e.g., $z = 0$ (female) and $z = 1$ (male), we may have $\mathbb{P}_{\mathsf{bias}}(Z = 0) > \mathbb{P}_{\mathsf{bias}}(Z = 1)$. For the purpose of promoting fair sample generation, we employ a balanced yet small reference dataset. As mentioned in Choi et al. (2020), one can obtain such balanced reference dataset without access to demographic labels (23&me 2016; Hong 2016; Yoshida 2014); see the supplementary for details. Let $\mathbb{P}_{\mathsf{ref}}$ be the distribution w.r.t. a reference sample $x_{\mathsf{ref}}^{(i)} \in \mathcal{X}$ where $\mathbb{P}_{\mathsf{ref}}(Z = 0) \approx \mathbb{P}_{\mathsf{ref}}(Z = 1)$. In practice, the number of the reference samples is often much smaller than that of training examples: $m_{\mathsf{ref}} \ll m_{\mathsf{bias}}$. Denote by $\hat{x} := G(w) \in \mathcal{X}$ the generated sample fed by a random noise input $w \in \mathcal{W}$. Let $\mathbb{P}_G$ and $\mathbb{P}_W$ be distributions w.r.t. the generated samples and the random noise input respectively.

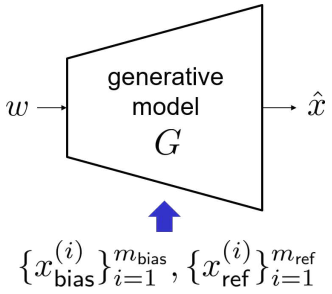

Figure 1: Part of a fair generative model that intends to yield generated samples with the equal size over demographic groups. We employ training data $\{x_{\mathsf{bias}}^{(i)}\}_{i=1}^{m_{\mathsf{bias}}}$ (potentially biased) and balanced reference data $\{x_{\mathsf{ref}}^{(i)}\}_{i=1}^{m_{\mathsf{ref}}}$. See Figure 2 for the entire structure of the proposed model. Here m_{bias} (or m_{ref}) denotes the number of training samples (or reference samples).

As a fairness measure that will be employed for the purpose of evaluating our framework to be presented in the next section, we consider *fairness discrepancy* proposed by Choi et al. (2020). It quantifies how $\mathbb{P}_G$ differs from $\mathbb{P}_{\mathsf{ref}}$ w.r.t. a certain sensitive attribute, formally defined below.

Definition 1 (Fairness Discrepancy (Choi et al. 2020))**.** *Fairness discrepancy between $\mathbb{P}_{\mathsf{ref}}$ and $\mathbb{P}_G$ w.r.t. a sensitive attribute $z \in \{z_1, \ldots, z_{|\mathcal{Z}|}\}$ is defined as:*

$$\mathcal{F}(\mathbb{P}_{\mathsf{ref}}, \mathbb{P}_G) := \|\mathbf{p}_{\mathsf{ref}}(z) - \mathbf{p}_G(z)\|_2 \tag{1}$$

where

$$\mathbf{p}_{\mathsf{ref}}(z) := [\mathbb{P}_{\mathsf{ref}}(Z = z_1) \ \cdots \ \mathbb{P}_{\mathsf{ref}}(Z = z_{|\mathcal{Z}|})]^{\mathsf{T}};$$
$$\mathbf{p}_G(z) := [\mathbb{P}_G(\hat{Z} = z_1) \ \cdots \ \mathbb{P}_G(\hat{Z} = z_{|\mathcal{Z}|})]^{\mathsf{T}}.$$

Here $\hat{Z}$ denotes the prediction of the sensitive attribute w.r.t. a generated sample. We assume that $\hat{Z}$ is available from a pre-trained attribute classifier. As in Choi et al.

(2020), the attribute classifier is employed only for the purpose of evaluation, and is trained based on another real dataset, e.g., like the one mentioned in Choi et al. (2020): the standard train and validation splits of CelebA. For faithful evaluation, we employ a reliable attribute classifier, which provides a sufficiently high accuracy, say 98%, for gender classification.

As a measure for the quality of generated samples that may compete with the fairness measure, we employ a well-known measure: Fréchet Inception Distance (FID) (Heusel et al. 2017). It is defined as the Fréchet distance (Fréchet 1957) (also known as the second-order Wasserstein distance (Wasserstein 1969)) between real and generated samples approximated via the Gaussian distribution. The lower FID, the more realistic and diverse the generated samples are. For a more precise measure that represents sample quality of each sensitive group, we consider FID computed *within* each demographic, called *intra* FID (Miyato and Koyama 2018; Zhang et al. 2019; Wang et al. 2020). Computing intra FID requires the knowledge on group identities of generated samples. Since demographic labels are not available in our setting, again we rely upon the attribute classifier (that we introduced above) for predicting demographic information of the generated samples.

GAN-based generative model: Our framework (to be presented soon) builds upon one powerful generative model: Generative Adversarial Networks (GANs) (Goodfellow et al. 2014). The GANs comprise two competing players: (i) discriminator $D(\cdot)$ that wishes to discriminate real samples against generated samples; and (ii) generator $G(\cdot)$ that intends to fool the discriminator by producing realistic generated samples. As our base framework, we consider a prominent f-GAN (Nowozin, Cseke, and Tomioka 2016), which subsumes many divergence-based GANs as a special case. For a convex function f satisfying $f(1) = 0$, the f-GAN optimization w.r.t. the training data distribution $\mathbb{P}_{\mathsf{bias}}$ reads:

$$\min_G \max_D \; \mathbb{E}_{\mathbb{P}_{\mathsf{bias}}}[D(X)] - \mathbb{E}_{\mathbb{P}_G}[f^*(D(X))] \tag{2}$$

where f^* is the Fenchel-conjugate of f: $f^*(t) = \sup_u\{ut - f(u)\}$. One can recover the original GAN optimization (Goodfellow et al. 2014) via $f(u) = u\log u - (u+1)\log(u+1) + 2\log 2$. One important property of the above optimization is that solving the inner problem induces the discriminator to learn the density ratio between training and generated data distributions (Song and Ermon 2020). It has been shown in Nowozin, Cseke, and Tomioka (2016) that plugging the optimal discriminator having such knowledge on the densities into (2) yields the following equivalent optimization:

$$\min_G \mathsf{D}_f(\mathbb{P}_{\mathsf{bias}}||\mathbb{P}_G) \tag{3}$$

where $\mathsf{D}_f(\mathbb{P}_{\mathsf{bias}}||\mathbb{P}_G)$ indicates the f-divergence between training and generated data distributions: $\mathsf{D}_f(\mathbb{P}_{\mathsf{bias}}||\mathbb{P}_G) := \sum_{x\in\mathcal{X}} \mathbb{P}_G(x) f\left(\mathbb{P}_{\mathsf{bias}}(x)/\mathbb{P}_G(x)\right)$. Notice that minimizing the f-divergence in (3) encourages generated samples to respect the biased distribution $\mathbb{P}_{\mathsf{bias}}$, aggravating the fairness performance. In the next section, we will present a generalized framework that intends to equip (3) with a fairness aspect.

Proposed Framework

Divergence-Based Fairness Regularization

One conventional approach to impose a fairness constraint is to incorporate a fairness-associated-regularization term. Here a natural question arises. What is a proper regularization term that can well capture the unfairness of a model?

Since we are interested in minimizing the fairness metric (*fairness discrepancy*, defined in (1)), one may want to directly use it as a regularization term. However, this approach is not permissible in our setting, as the computation of fairness discrepancy requires the knowledge of demographic labels which are assumed to be unavailable.

Here we propose a different regularization term for fairness which is tractable to compute. Remember our framework employs a balanced reference dataset with $\mathbb{P}_{\mathsf{ref}}$. For fairness, we want to make $\mathbb{P}_G$ as similar as possible to $\mathbb{P}_{\mathsf{ref}}$, so a divergence measure between $\mathbb{P}_G$ and $\mathbb{P}_{\mathsf{ref}}$ can serve to quantify the degree of unfairness. Taking such measure as a regularization term, we obtain:

$$\min_G \; (1-\lambda)\cdot \mathsf{D}_f(\mathbb{P}_{\mathsf{bias}}||\mathbb{P}_G) + \lambda \cdot \mathsf{D}_{\mathsf{fair}}(\mathbb{P}_{\mathsf{ref}}||\mathbb{P}_G) \tag{4}$$

where $\mathsf{D}_{\mathsf{fair}}(\mathbb{P}_{\mathsf{ref}}||\mathbb{P}_G)$ indicates a divergence measure for fairness (subject to our choice), and $\lambda \in [0,1]$ denotes a normalized regularization factor that balances the sample quality against the fairness constraint. Notice that the regularization term introduced above is indeed computable, e.g., via empirical versions of $\mathbb{P}_{\mathsf{ref}}$ and $\mathbb{P}_G$ constructed from given data samples.

Robust Regularization via LeCam Divergence

One challenge that arises in (4) is that in practice, the size of reference dataset is often very small relative to training dataset, i.e., $m_{\mathsf{ref}} \ll m_{\mathsf{bias}}$, so measuring $\mathsf{D}_{\mathsf{fair}}(\mathbb{P}_{\mathsf{ref}}||\mathbb{P}_G)$ may often be highly inaccurate, thereby degrading the fairness performance.

We address this challenge by employing one divergence measure in the f-divergence family, LeCam (LC)-divergence (a.k.a. triangular discrimination) (Le Cam 2012). This choice is inspired by a recent study (Tseng et al. 2021) in which a properly-weighted version of LC-divergence is shown to be robust to the size of data, thus GAN training based on such weighted divergence yields little performance degradation in small-sized datasets. Replacing the existing regularization term in (4) with the weighted LC-divergence, we get:

$$\min_G \; (1-\lambda)\cdot \mathsf{D}_f(\mathbb{P}_{\mathsf{bias}}||\mathbb{P}_G) + \lambda \cdot \mu\mathsf{D}_{\Delta}(\mathbb{P}_{\mathsf{ref}}||\mathbb{P}_G) \tag{5}$$

where μ denotes a non-negative weight (another hyperparameter of which the role will be explained in detail shortly), and $\mathsf{D}_{\Delta}(\mathbb{P}_{\mathsf{ref}}||\mathbb{P}_G)$ indicates the LC-divergence between $\mathbb{P}_{\mathsf{ref}}$ and $\mathbb{P}_G$:

$$\mathsf{D}_{\Delta}(\mathbb{P}_{\mathsf{ref}}||\mathbb{P}_G) := \sum_{x\in\mathcal{X}} \frac{(\mathbb{P}_{\mathsf{ref}}(x) - \mathbb{P}_G(x))^2}{\mathbb{P}_{\mathsf{ref}}(x) + \mathbb{P}_G(x)}.$$

Now the question is how to express $\mu\mathsf{D}_{\Delta}(\mathbb{P}_{\mathsf{ref}}||\mathbb{P}_G)$ in terms of an optimization variable G. To this end, we invoke a variational optimization technique (Tseng et al. 2021)

that allows us to translate $\mu \mathsf{D}_\Delta(\mathbb{P}_{\mathsf{ref}}||\mathbb{P}_G)$ into a function optimization with a regularized objective. In our framework, the existing discriminator function D is dedicated for expressing $\mathsf{D}_f(\mathbb{P}_{\mathsf{bias}}||\mathbb{P}_G)$, so we introduce another discriminator function D_{ref} for expressing $\mu \mathsf{D}_\Delta(\mathbb{P}_{\mathsf{ref}}||\mathbb{P}_G)$. Employing the translation technique (Tseng et al. 2021), one can show that (5) is equivalent to the following nested optimization (see the proof of Proposition 1 below for derivation):

$$\begin{aligned}
&\max_{D} \ \mathbb{E}_{\mathbb{P}_{\mathsf{bias}}}[D(X)] - \mathbb{E}_{\mathbb{P}_G}[f^*(D(X))] \\
&\max_{D_{\mathsf{ref}}} \ \mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}[D_{\mathsf{ref}}(X)] - \mathbb{E}_{\mathbb{P}_G}[D_{\mathsf{ref}}(X)] - \frac{1}{2(\mu+\alpha)} R_\Delta \\
&\min_{G} \ -(1-\lambda)\, \mathbb{E}_{\mathbb{P}_G}[f^*(D(X))] - \lambda\, \mathbb{E}_{\mathbb{P}_G}[D_{\mathsf{ref}}(X)]
\end{aligned} \tag{6}$$

where α denotes an exponential moving average of D_{ref} w.r.t. reference samples (see the supplementary for details), and R_Δ indicates a regularization term for D_{ref} defined as:

$$R_\Delta := \mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}\big[\|D_{\mathsf{ref}}(X)+\alpha\|^2\big] + \mathbb{E}_{\mathbb{P}_G}\big[\|D_{\mathsf{ref}}(X)-\alpha\|^2\big].$$

We see that the hyperparameter μ serves as a regularization factor (together with α) in the D_{ref} optimization. Notice that the optimization for D is the same as that in (2), which serves to implement $\mathsf{D}_f(\mathbb{P}_{\mathsf{bias}}||\mathbb{P}_G)$. The new regularized optimization w.r.t. D_{ref} together with the second term in the generator objective yields $\mu \mathsf{D}_\Delta(\mathbb{P}_{\mathsf{ref}}||\mathbb{P}_G)$. The translated three-player optimization can then be implemented. For instance, we parameterize (D, D_{ref}, G) with three neural networks and then employ three-way alternating gradient descent (Goodfellow et al. 2014) for the parameterized neural networks; see Algorithm 1 in the supplementary for details. The equivalence between (5) and the translated optimization (6) can be readily shown via the proof technique used in Tseng et al. (2021). See Proposition 1 below for the formal statement of the equivalence and the proof.

Proposition 1. *Consider a three-player optimization in* (6)*. Assume that given a fixed generator G, an exponential moving average of D_{ref} w.r.t. reference data samples converges to a stationary value $\alpha > 0$. Then, under the optimal discriminators D^* and D^*_{ref}, the optimization for generator G is equivalent to* (5)*.*

Proof. Since $\mathbb{E}_{\mathbb{P}_{\mathsf{bias}}}[D(X)]$ and $\mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}[D_{\mathsf{ref}}(X)]$ are irrelevant to G, the optimization for G in (6) can be written as:

$$\begin{aligned}
&\min_{G} \ -(1-\lambda)\, \mathbb{E}_{\mathbb{P}_G}[f^*(D(X))] - \lambda\, \mathbb{E}_{\mathbb{P}_G}[D_{\mathsf{ref}}(X)] \\
&= \min_{G} \ (1-\lambda)\big\{\mathbb{E}_{\mathbb{P}_{\mathsf{bias}}}[D(X)] - \mathbb{E}_{\mathbb{P}_G}[f^*(D(X))]\big\} \\
&\qquad + \lambda\big\{\mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}[D_{\mathsf{ref}}(X)] - \mathbb{E}_{\mathbb{P}_G}[D_{\mathsf{ref}}(X)]\big\}.
\end{aligned} \tag{7}$$

The optimization for D in (6) is the same as that in (2), so it yields the same optimal function D^* as (2). Plugging D^* into the terms associated with $(1-\lambda)$ in (7), we obtain the first term in the desired formula (5): $(1-\lambda)\cdot \mathsf{D}_f(\mathbb{P}_{\mathsf{bias}}||\mathbb{P}_G)$. For D_{ref}, the optimization in (6) yields the following (Tseng et al. 2021):

$$D^*_{\mathsf{ref}}(x) = \mu \cdot \frac{\mathbb{P}_{\mathsf{ref}}(x) - \mathbb{P}_G(x)}{\mathbb{P}_{\mathsf{ref}}(x) + \mathbb{P}_G(x)}.$$

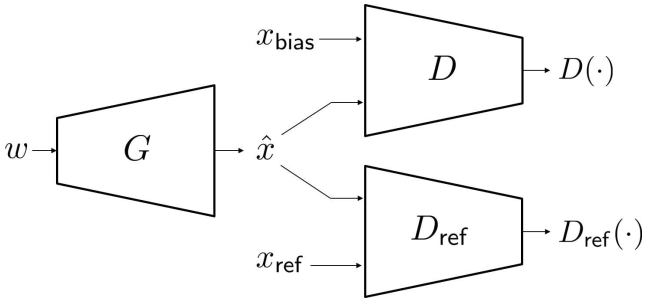

Figure 2: The architecture of the proposed three-player optimization, reflected in (6).

Plugging D^*_{ref} into the terms associated with λ in (7), we get the desired LeCam divergence term (Tseng et al. 2021). This completes the proof. □

For the assumption w.r.t. the convergence of an exponential moving average, we found the same justification in Tseng et al. (2021) can carry over our framework. More precisely, we empirically observed that the output values of D_{ref} w.r.t. reference samples converge to a stationary point in most of our experiments. In some rare cases where such convergence may not be naturally attained, one may artificially *implement* the convergence by making the moving average stop updating after a sufficient number of training iterations (Tseng et al. 2021).

Remark 1 (How to implement LC-divergence?)**.** *For implementing the LC-divergence, we adopt the same methods as Tseng et al. (2021) which we found empirically more beneficial in our framework. As described in Algorithm 1 (in the supplementary), we apply the hinge loss (Lim and Ye 2017; Tran, Ranganath, and Blei 2017) for D_{ref}, which encourages D_{ref} to constrain its output value, thus improving the training stability. Also, we introduce an additional moving average w.r.t. generated samples, say α_G, and incorporate it in R_Δ: $R_\Delta = \mathbb{E}_{\mathbb{P}_{\mathsf{ref}}}[\|D_{\mathsf{ref}}(X)-\alpha_G\|^2]+\mathbb{E}_{\mathbb{P}_G}[\|D_{\mathsf{ref}}(X)-\alpha\|^2]$. We found that the use of the two moving averages offers greater performances relative to the single moving-average counterpart. In addition, we found fixing the weight of R_Δ enables more stabilized training, yielding the performance gain compared to the one with a variable weight in* (6)*, i.e., $-1/\{2(\mu+\alpha)\}$.*

Remark 2 (Three-way battles)**.** *Figure 2 illustrates the entire architecture of the translated three-level optimization. Here we see interesting* three-way battles*. The first is a well-known battle between the generator G and the 1st discriminator D. Remember D^* acts as a density ratio estimator between $\mathbb{P}_{\mathsf{bias}}$ and $\mathbb{P}_G$. So one can interpret D^* as the strength of distinguishing real (potentially biased) samples against generated samples. On the other hand, the generator intends to fool D, thus promoting realistic samples. The second battle is in between the generator and the 2nd discriminator D_{ref}. The same interpretation can be made from $D^*_{\mathsf{ref}}(x) = \mu \cdot \frac{\mathbb{P}_{\mathsf{ref}}(x)-\mathbb{P}_G(x)}{\mathbb{P}_{\mathsf{ref}}(x)+\mathbb{P}_G(x)}$ (the ability to distinguish balanced reference samples against the generated samples).*

This way, the generator G is encouraged to produce balanced yet less realistic *(due to the small-sized reference set) samples, thus pitting the 1st discriminator against the 2nd discriminator* indirectly. *The last battle is in between the 1st and 2nd discriminators. This tension is* directly *controlled by the fairness tuning knob λ; see corresponding tradeoff curves presented in Figure 4 in the next section. It turns out the three-way tradeoff relationships established via our LC-based framework are greatly balanced, thus achieving significant performances both in fairness and sample quality. This is empirically demonstrated in the next section; see Table 1 for details.*

Experiments

We conduct experiments on three benchmark real datasets: CelebA (Liu et al. 2015), UTKFace (Zhang, Song, and Qi 2017), and FairFace (Karkkainen and Joo 2021). We implement our algorithm in PyTorch (Paszke et al. 2019), and all experiments are performed on servers with TITAN RTX and Quadro RTX 8000 GPUs. For our algorithm, all the simulation results (to be reported) are the ones averaged over five trials with distinct seeds in training.

Setup

Datasets: Our construction of $\mathcal{D}_{\mathsf{bias}}$ and $\mathcal{D}_{\mathsf{ref}}$ respects the method described in Choi et al. (2020). Only for the purpose of data construction, we have an access to sensitive attributes z, so as to control the ratio of demographic group sizes. For CelebA, we consider two scenarios depending on the number of focused attributes: (i) CelebA-single (gender); (ii) CelebA-multi (two attributes: gender and hair color). Training data $\mathcal{D}_{\mathsf{bias}}$ is constructed to have $9:1$ ratio (female vs. male) samples where $m_{\mathsf{bias}} = 67507$. We take balanced samples for $\mathcal{D}_{\mathsf{ref}}$ ($1:1$ ratio). For CelebA-multi, we have four groups: (i) (female, non-black); (ii) (male, non-black); (iii) (female, black); (iv) (male, black). For $\mathcal{D}_{\mathsf{bias}}$, we take $85:15$ ratio samples (non-black hair vs. black hair) where $m_{\mathsf{bias}} = 60000$. For UTKFace dataset, we consider a race attribute: white vs. non-white. We take $9:1$ ratio biased samples with $m_{\mathsf{bias}} \approx 10000$. For FairFace dataset, we consider another type of race categorized as white vs. black. We also take the $9:1$ ratio biased samples yet with $m_{\mathsf{bias}} \approx 20000$. Additionally, we consider different settings of sensitive attributes and bias ratio that are more difficult to work on; see the supplementary for details. A wide range of the reference set size is taken into consideration. We focus mainly on two sizes: (i) 10% ($m_{\mathsf{ref}} \approx 0.1 m_{\mathsf{bias}}$); (ii) 25% ($m_{\mathsf{ref}} \approx 0.25 m_{\mathsf{bias}}$). To demonstrate the robustness of our proposed approach to the reference set size, we also consider small sizes of the reference set all the way down to 1%. See the supplementary for more details.

Baselines: We consider three baselines. The first baseline, say Baseline I, is a *non-fair* algorithm trained on the aggregated dataset $\mathcal{D}_{\mathsf{bias}} \cup \mathcal{D}_{\mathsf{ref}}$. The second baseline, say Baseline II, is the same non-fair algorithm yet trained only with a small balanced reference set $\mathcal{D}_{\mathsf{ref}}$. The last is the state of the art, Choi et al. (2020). For all three baselines, we employ the hinge loss optimization (Lim and Ye 2017; Tran, Ranganath, and Blei 2017).

Attribute classifiers: As mentioned in the second section (near Definition 1), we employ attribute classifiers, only for the purpose of evaluating our twin measures: (i) fairness discrepancy (defined in (1)); (ii) intra FID. We introduce four different attribute classifiers for predicting sensitive attributes in the following scenarios: (i) gender for CelebA-single; (ii) gender and hair-color for CelebA-multi; (iii) white-vs-non-white race for UTKFace; (iv) white-vs-black race for FairFace. For all the classifiers, we use a variant of ResNet18 (He et al. 2016). CelebA and FairFace clas-

Reference set size		25%	10%	5%	2.5%	1%
Baseline I	Intra FID	$\underline{12.00 \pm 0.069}$	$\mathbf{12.73 \pm 0.053}$	$\mathbf{13.54 \pm 0.074}$	$\mathbf{13.79 \pm 0.072}$	$\mathbf{15.89 \pm 0.094}$
	Fairness	0.495 ± 0.001	0.554 ± 0.002	0.559 ± 0.001	0.566 ± 0.002	0.576 ± 0.002
Baseline II	Intra FID	23.81 ± 0.118	32.31 ± 0.109	40.07 ± 0.062	67.70 ± 0.112	92.34 ± 0.131
	Fairness	0.093 ± 0.002	0.115 ± 0.002	$\underline{0.120 \pm 0.003}$	$\underline{0.150 \pm 0.003}$	0.455 ± 0.002
Choi et al. (2020)	Intra FID	20.68 ± 0.076	25.74 ± 0.079	30.15 ± 0.037	30.40 ± 0.041	31.49 ± 0.074
	Fairness	$\underline{0.065 \pm 0.002}$	$\underline{0.104 \pm 0.002}$	0.126 ± 0.001	0.237 ± 0.003	$\underline{0.344 \pm 0.002}$
Proposed	Intra FID	$\mathbf{11.48 \pm 0.814}$	$\underline{14.50 \pm 0.996}$	$\underline{14.64 \pm 0.626}$	$\underline{17.16 \pm 1.607}$	$\underline{23.11 \pm 0.797}$
	Fairness	$\mathbf{0.037 \pm 0.007}$	$\mathbf{0.039 \pm 0.013}$	$\mathbf{0.118 \pm 0.007}$	$\mathbf{0.129 \pm 0.010}$	$\mathbf{0.146 \pm 0.022}$

Table 1: Performance comparison on CelebA dataset. We provide the results for CelebA-single in which a *single* attribute (gender) is employed. See the supplementary for the results on CelebA-multi concerning two attributes: gender and hair color. Baseline I is a non-fair algorithm with the hinge loss (Lim and Ye 2017; Tran, Ranganath, and Blei 2017), and trained with the aggregated data $\mathcal{D}_{\mathsf{bias}} \cup \mathcal{D}_{\mathsf{ref}}$. Baseline II is the same non-fair algorithm yet trained only with the small yet balanced reference dataset $\mathcal{D}_{\mathsf{ref}}$. Choi et al. (2020) is the state of the art. "Intra FID" refers to Fréchet Inception Distance (Heusel et al. 2017) computed within each group (Miyato and Koyama 2018; Zhang et al. 2019). We provide Intra FIDs for the minority group (i.e., male) herein and leave results for the majority group in the supplementary. The lower intra FID, the more realistic and diverse samples. "Fairness" is *fairness discrepancy* introduced by Choi et al. (2020); see (1) for the definition. The lower, the fairer. For each measure, we mark the best result in bold and the second-best with underline. The reference set size indicates a ratio relative to training data.

sifiers are trained over the standard train and validation splits of CelebA and FairFace, respectively. For training the UTKFace classifier, we use $8:1:1$ splits of UTKFace dataset. We found that our evaluation is often sensitive to the performances of the attribute classifiers; see the supplementary for a detailed discussion.

Hyperparameter search: For implementation of all three baselines (Baseline I, Baseline II, and Choi et al. (2020)) and the proposed framework, we all employ the BigGAN architecture (Brock, Donahue, and Simonyan 2019). In other words, we parameterize G, D, and D_{ref} with the neural-net architecture introduced in Brock, Donahue, and Simonyan (2019). We leave details in the supplementary. We also conduct a complexity analysis of our algorithm with a comparison to the state of the art (Choi et al. 2020); see the supplementary for details.

Results

Table 1 provides performance comparison with the three baselines on CelebA dataset. For a wide range of the reference set size, our approach outperforms the state of the art (Choi et al. 2020) both in fairness ("Fairness discrep-

		UTKFace		FairFace	
Reference set size		25%	10%	25%	10%
Baseline I	Intra FID	**18.86 ± 0.117**	**19.89 ± 0.119**	**22.96 ± 0.047**	**25.76 ± 0.068**
	Fairness	0.400 ± 0.003	0.453 ± 0.002	0.386 ± 0.003	0.434 ± 0.002
Baseline II	Intra FID	35.73 ± 0.077	83.51 ± 0.071	45.20 ± 0.055	83.76 ± 0.118
	Fairness	**0.007 ± 0.003**	**0.010 ± 0.003**	**0.009 ± 0.002**	0.105 ± 0.002
Choi et al. (2020)	Intra FID	35.04 ± 0.103	36.43 ± 0.231	32.82 ± 0.073	33.33 ± 0.076
	Fairness	0.178 ± 0.003	0.285 ± 0.003	0.213 ± 0.002	0.317 ± 0.002
Proposed	Intra FID	20.62 ± 1.294	27.24 ± 4.125	24.24 ± 0.228	30.76 ± 2.072
	Fairness	0.072 ± 0.010	0.091 ± 0.022	0.078 ± 0.005	**0.094 ± 0.014**

Table 2: Performance comparison on UTKFace and FairFace datasets. All the settings and baselines are the same as those in Table 1, except for different datasets. For each measure, we mark the best result in bold and the second-best with underline.

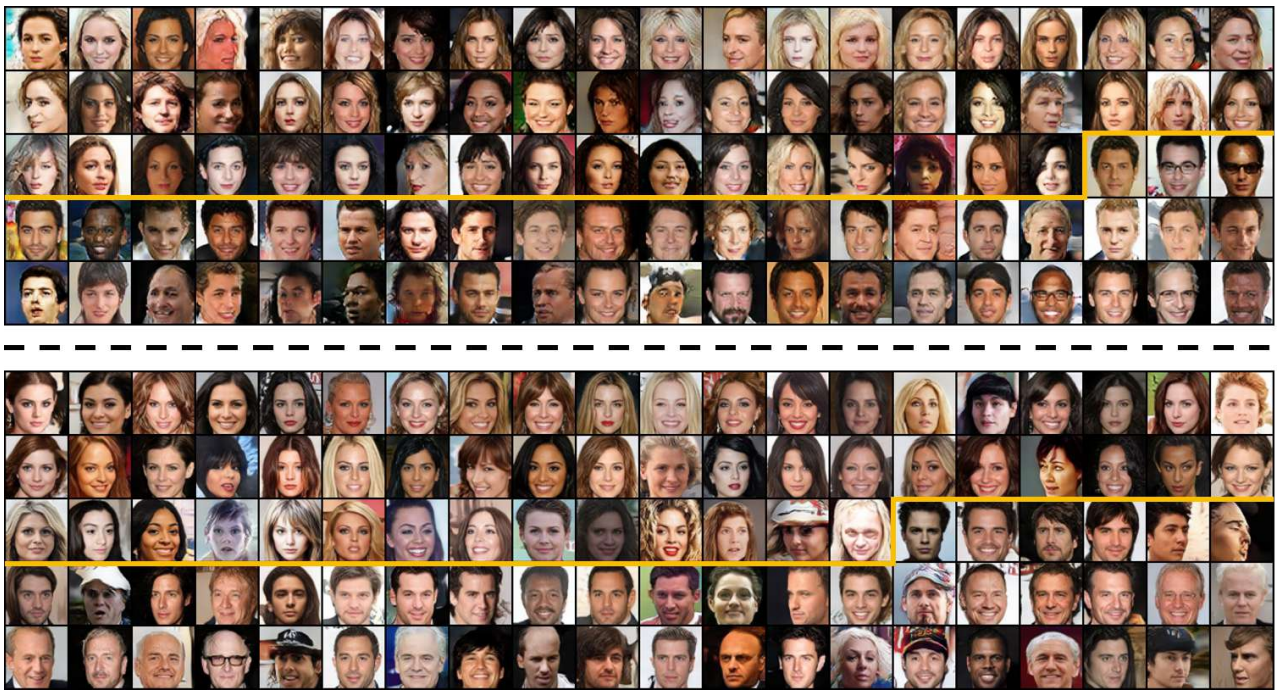

Figure 3: (Top) Generated samples by Choi et al. (2020) trained on CelebA-single with 10% reference set size. Faces above the yellow line are female (57 pictures), while the rest are male (43). Intra FIDs are around 21.07 (female) and 25.74 (male); (Bottom) Generated samples by the proposed approach under the same setting. We obtain 54 females and 46 males, yet producing more realistic sample images, reflected in much lower intra FIDs, around 9.31 (female) and 14.50 (male).

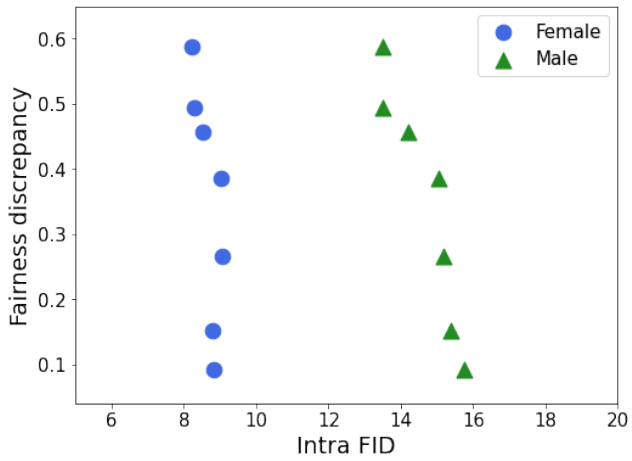

Figure 4: Fairness-quality tradeoff curves evaluated on CelebA-single with a 10%-sized reference set. Each point is obtained with a particular λ, the fairness tuning knob in our framework. Blue dot points indicate performances for female group, and green triangles are for male group.

	Intra FID		
	Female	Male	Fairness
JS	$\underline{12.80 \pm 1.499}$	$\underline{17.83 \pm 1.173}$	0.087 ± 0.012
KL	15.24 ± 0.371	22.47 ± 0.331	0.077 ± 0.019
χ^2	16.01 ± 1.601	25.25 ± 1.877	0.058 ± 0.019
W	16.51 ± 1.244	24.54 ± 2.731	$\underline{0.047 \pm 0.033}$
LC	$\mathbf{9.31 \pm 0.825}$	$\mathbf{14.50 \pm 0.996}$	$\mathbf{0.039 \pm 0.013}$

Table 3: Performance comparison with other fairness regularizers on CelebA-single with the 10% reference set size. "JS" refers to a regularization method based on Jensen-Shannon divergence implemented via Goodfellow et al. (2014). "KL" is the one built upon Kullback-Leibler divergence (Nowozin, Cseke, and Tomioka 2016). "χ^2" represents the one implementing Pearson-χ^2 divergence (Mao et al. 2017). "W" refers to a regularization with Wasserstein distance (Gulrajani et al. 2017). "Female" (or "Male") refers to intra FID for female (or male) group. For each measure, we mark the best result in bold and the second-best with underline.

ancy") and sample quality ("Intra FID"). The lower, the better for all the measures. Notice even for the 1% reference set size, our algorithm offers still respectable fairness performance. This corroborates the robustness aspect of the LC-divergence to the small reference dataset, which we employ for encouraging fair sample generation. On the other hand, Choi et al. (2020) suffers from fairness degradation starting from 2.5%, exhibiting its sensitivity to the size of the reference data.

Table 2 concerns UTKFace and FairFace datasets. We consider the same settings as in Table 1. One significant distinction w.r.t. CelebA dataset is that training and reference set sizes are much smaller; see the supplementary for more details. Hence, as expected, the overall performances are worse than those on CelebA. Even in this small data regime, we observe the same trends on the performance benefits of ours relative to the baselines. Refer to the supplementary for intra FIDs w.r.t. the majority groups.

Figure 3 visualizes generated samples on CelebA-single with the 10% reference set size. The top figure corresponds to Choi et al. (2020), while the bottom is due to the proposed algorithm. For each figure, faces above the yellow lines are female samples, while the rest are male samples. Here the gender is predicted via the attribute classifier with around 98% accuracy. While both approaches yield well-balanced samples ($57 : 43$ for Choi et al. (2020), and $54 : 46$ for ours), our algorithm produces more realistic sample images. This is reflected in lower intra FIDs, around 9.31 (female group) and 14.50 (male group). On the other hand, Choi et al. (2020) offers intra FIDs of around 21.07 and 25.74 for female and male groups, respectively. See the supplementary for the intra FID values for both groups. In the supplementary, we also provide generated samples for UTKFace and FairFace datasets.

Figure 4 demonstrates tradeoff curves between fairness and sample quality offered by our framework. Each point in the curves corresponds to performance with a specific λ value in $\{0, 0.3, 0.4, 0.45, 0.5, 0.63, 0.7\}$. Observe that as the fairness tuning knob λ increases, fairness performance gets improved (having lower fairness discrepancy) at the expense of the degraded sample quality, reflected in larger intra FID. This validates the role of λ as a tuning knob that controls the strength of fairness.

Table 3 provides performance comparison with other fairness regularizations that employ different divergence measures. Observe that among the considered regularization methods, our LC-based approach offers the best performances both in fairness and sample quality. It also yields the smallest discrepancy between intra FIDs of different groups. Another noticeable observation is that our divergence-based regularization approach outperforms Choi et al. (2020) for a variety of other divergence measures not limited to the LC-divergence; also see Table 1 for detailed comparison.

Conclusion

We introduced an LC-based optimization framework for a fair generative model that well tradeoffs the fairness performance (quantified as fairness discrepancy) against sample quality (reflected in intra FID). Inspired by the equivalence between the LC-divergence and function optimization, we also developed an equivalent three-player optimization which can readily be implemented via neural-net parameterization. Our algorithm offers better performances than the state of the art both in fairness and sample quality, exhibiting more significant performances particularly for practically-relevant scenarios where the access to balanced dataset is limited. One future work of interest is to push forward for more challenging scenarios where the reference dataset is not available.

Acknowledgments

This work was supported by Field-oriented Technology Development Project for Customs Administration through National Research Foundation of Korea (NRF) funded by the Ministry of Science & ICT and Korea Customs Service (No. NRF-2021M3I1A1097938) and by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)).

References

23&me. 2016. The real issue: Diversity in genetics research. https://blog.23andme.com/ancestry-reports/the-real-issue-diversity-in-genetics-research/. Accessed: 2023-03-20.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. *arXiv preprint arXiv:1701.07875*.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*.

Byrd, J.; and Lipton, Z. 2019. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*. PMLR.

Chang, C.-T.; Chuang, S.-P.; and Lee, H.-Y. 2018. Code-switching sentence generation by generative adversarial networks and its application to data augmentation. *arXiv preprint arXiv:1811.02356*.

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 (ICML)*. PMLR.

Fréchet, M. 1957. Sur la distance de deux lois de probabilité. *C. R. Acad. Sci. Paris*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27 (NeurIPS)*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. 2017. Improved training of wasserstein GANs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*.

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*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*.

Hong, E. 2016. 23andme has a problem when it comes to ancestry reports for people of color. https://qz.com/765879/23andme-has-a-race-problem-when-it-comes-to-ancestry-reports-for-non-whites/. Accessed: 2023-03-20.

Jalal, A.; Karmalkar, S.; Hoffmann, J.; Dimakis, A.; and Price, E. 2021. Fairness for Image Generation with Uncertain Sensitive Attributes. In *International Conference on Machine Learning*. PMLR.

Jang, T.; Zheng, F.; and Wang, X. 2021. Constructing a Fair Classifier with Generated Fair Data. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*.

Karkkainen, K.; and Joo, J. 2021. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kyono, T.; van Breugel, B.; Berrevoets, J.; and van der Schaar, M. 2021. DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks. *NeurIPS. cc*.

Le Cam, L. 2012. *Asymptotic methods in statistical decision theory*. Springer Science & Business Media.

Lee, J.; Kim, H.; Hong, Y.; and Chung, H. W. 2021. Self-Diagnosing GAN: Diagnosing Underrepresented Samples in Generative Adversarial Networks. *arXiv preprint arXiv:2102.12033*.

Lim, J. H.; and Ye, J. C. 2017. Geometric GAN. *arXiv preprint arXiv:1705.02894*.

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)*.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Paul Smolley, S. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Miyato, T.; and Koyama, M. 2018. cGANs with projection discriminator. *arXiv preprint arXiv:1802.05637*.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*.

Odena, A.; Olah, C.; and Shlens, J. 2017. Conditional image synthesis with auxiliary classifier gans. In *International conference on machine learning*. PMLR.

Ozawa, N. 2015. A functional analysis proof of Gromov's polynomial growth theorem. *arXiv preprint arXiv:1510.04223*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32 (NeurIPS)*.

Ren, M.; Zeng, W.; Yang, B.; and Urtasun, R. 2018. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*. PMLR.

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*.

Shu, K.; Zheng, G.; Li, Y.; Mukherjee, S.; Awadallah, A. H.; Ruston, S.; and Liu, H. 2020. Leveraging multi-source weak social supervision for early detection of fake news. *arXiv preprint arXiv:2004.01732*.

Song, J.; and Ermon, S. 2020. Bridging the gap between f-gans and wasserstein gans. In *International Conference on Machine Learning*. PMLR.

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

Tran, D.; Ranganath, R.; and Blei, D. M. 2017. Deep and hierarchical implicit models. *arXiv preprint arXiv::1702.08896*.

Tseng, H.-Y.; Jiang, L.; Liu, C.; Yang, M.-H.; and Yang, W. 2021. Regularizing Generative Adversarial Networks under Limited Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wang, X.; Wang, K.; and Lian, S. 2019. A survey on face data augmentation. *arXiv preprint arXiv:1904.11685*.

Wang, Y.; Chen, Y.-C.; Zhang, X.; Sun, J.; and Jia, J. 2020. Attentive normalization for conditional image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wasserstein, L. N. 1969. Markov processes over denumerable products of spaces describing large systems of automata. *Probl. Inform. Transmission*.

Xu, D.; Wu, Y.; Yuan, S.; Zhang, L.; and Wu, X. 2019a. Achieving causal fairness through generative adversarial networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*.

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)*. IEEE.

Xu, D.; Yuan, S.; Zhang, L.; and Wu, X. 2019b. Fairgan+: Achieving fair data generation and classification through generative adversarial nets. In *2019 IEEE International Conference on Big Data (Big Data)*. IEEE.

Yehudayoff, A. 2020. Pointer chasing via triangular discrimination. *Combinatorics, Probability and Computing*, 29(4): 485–494.

Yi, X.; Walia, E.; and Babyn, P. 2019. Generative adversarial network in medical imaging: A review. *Medical image analysis*.

Yoshida, N. 2014. Revolutionizing Data Collection: From "Big Data" to "All Data". https://blogs.worldbank.org/developmenttalk/revolutionizing-data-collection-big-data-all-data. Accessed: 2023-03-20.

Yu, N.; Li, K.; Zhou, P.; Malik, J.; Davis, L.; and Fritz, M. 2020. Inclusive gan: Improving data and minority coverage in generative models. In *European Conference on Computer Vision*. Springer.

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2019. Self-attention generative adversarial networks. In *International conference on machine learning*. PMLR.

Zhang, Z.; Song, Y.; and Qi, H. 2017. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.