ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230484

Fair Latent Deep Generative Models (FLDGMs) for Syntax-Agnostic and Fair Synthetic Data Generation

Resmi Ramachandranpillai[a;*], **Md Fahim Sikder**[*a] **and Fredrik Heintz**[a]

[a]Department of Computer and Information Science (IDA), Linköping University, Sweden
ORCiD ID: Resmi Ramachandranpillai https://orcid.org/0000-0002-4302-9327

Abstract. Deep Generative Models (DGMs) for generating synthetic data with properties such as quality, diversity, fidelity, and privacy is an important research topic. Fairness is one particular aspect that has not received the attention it deserves. One difficulty is training DGMs with an in-process fairness objective, which can disturb the global convergence characteristics. To address this, we propose Fair Latent Deep Generative Models (FLDGMs) as enablers for more flexible and stable training of fair DGMs, by first learning a syntax-agnostic, model-agnostic fair latent representation (low dimensional) of the data. This separates the fairness optimization and data generation processes thereby boosting stability and optimization performance. Moreover, data generation in the low dimensional space enhances the accessibility of models by reducing computational demands. We conduct extensive experiments on image and tabular domains using Generative Adversarial Networks (GANs) and Diffusion Models (DMs) and compare them to the state-of-the-art in terms of fairness and utility. Our proposed FLDGMs achieve superior performance in generating high-quality, high-fidelity, and high-diversity fair synthetic data compared to the state-of-the-art fair generative models.

1 Introduction

Deep Generative Models (DGMs) have achieved substantial progress in learning to approximate the real data distribution as closely as possible. In particular, Generative Adversarial Networks (GANs) [10] and Diffusion Models (DMs) [15] are the most successful among the generative models for generating high-dimensional data. Existing generation methods based on GANs and DMs have focused on properties such as fidelity, quality, diversity, and privacy. Fidelity and quality relate to how closely synthetic data captures the distribution of real data. Diversity measures how successful they are in generating new distributions that are covered by real data. Lastly, privacy guarantees that synthetic data is not just a replication of real data, which is very important in sensitive domains [37].

Synthetic data fairness - generating fair data from biased data - is a much less-explored concept in the context of generative models. A few solutions to this problem such as FairGAN [38] and DECAF [33] have been proposed based on GANs to ensure fairness in the downstream tasks. A recent study shows that existing GAN techniques amplify the bias present in the training data resulting in more biased data in the target [14] including differential privacy generation schemes. Therefore, protected groups or people with certain sensitive or protected characteristics like ethnicity, gender, or religion [4], can have biased treatments in downstream models. As a result, safeguarding against discrimination or unfavorable outcomes a person's protected qualities [25] has become more crucial in ML.

Motivation. Altering training with an in-process fairness objective may disturb the quality-fairness tradeoff in Data Generation Process (DGP). Additionally, state-of-the-art fair generative models [38, 33] operate on pixel (for image) or attributes (for categorical and continuous features in tabular) and thus are highly dependent on the underlying datatypes (or syntax). Modeling these features in the high-dimensional space requires complex model architectures compatible with the underlying data types. This demands high computational resources for the generative models, which reduces the accessibility of these models to the general research community.

Research gap. There is a lack of study in learning fair DGMs to reach an optimal point between accessibility, fairness, quality, and flexibility (fine-tuning to various architectures, tasks, and fairness measures).

To this end, we propose Fair Latent Deep Generative Models (FLDGMs), both for GANs and DMs. Our FLDGMs are syntax-agnostic and stable and operate on low-dimensional continuous latent space. First, our approach starts with learning a fair compression using Variational AutoEncoders that enable fast sampling from the input domain and encourage quality in the target as the DGMs in the subsequent stage can focus on optimizing this compressed dimension. Second, the fair latent vectors can be used for various generative models (such as many versions of GANs and DMs) and applications independent of data-specific architectures, which makes the approach more generalizable. Third, it can be extended to impose various fairness constraints in the synthetic data given the pre-trained Variational AutoEncoders for the corresponding fairness measures, boosting flexibility. Fourth, since the generation is performed by either GANS or DMs, it can produce high-quality samples which contradict the approach described in [20]. Finally, the transformation from the generated fair latent space to the fair data space can be done in a single pass. To the best of our knowledge, there are no studies involving the capabilities mentioned above in fair data generation schemes.

Contributions. Our key contributions are four-fold: (i) We propose a novel formulation of a fair latent generative framework common to both GANs and Diffusion models; (ii) In contrast to previous works [38, 39, 21] which generate both fair and accurate synthetic data simultaneously, FLDGMs do not require a delicate weighting factor of generation quality and fairness penalty. Therefore, our approach

[*] Equal contribution
Corresponding author: resmi.ramachandran.pillai@liu.se; resmiramachandranpillai@gmail.com

Please check ArXiv or contact the authors for any appendices or supplementary material mentioned in the paper.

requires zero regularization of the latent space and ensures high-fidelity reconstructions and guarantees global convergence; (iii) The FLDGMs can be generalizable to data of any category, reducing the data pre-processing and modeling overhead in DGMs; (iv) Lastly, we conducted extensive experiments on tabular and image domains for various generative frameworks and compared the performance to the state-of-the-art in terms of fairness and data utility. Moreover, we also analyze the fidelity, diversity, and authenticity [1] of our proposed FLDGMs. The code and supplementary materials can be found at https://github.com/fahim-sikder/FLDGM.

2 Preliminaries

2.1 Algorithmic Fairness

This section defines disparate treatment and disparate impact measures of algorithmic fairness. Given a biased dataset $\mathbf{D} = \{X, S, Y\}$, where $X \in \mathcal{X}$, $S \in \mathcal{S}$, and $Y \in \mathcal{Y}$ respectively denote the set of non-sensitive, sensitive and target attributes. The features S and Y are categorical.

Definition 1: Fairness Through Unawareness (FTU) [11] - Let h be a prediction function, $h : X \rightarrow \hat{Y}$, and $\hat{Y}$ be the prediction outcome. The function h satisfies FTU if the sensitive attributes S are not explicitly used by h to obtain $\hat{Y}$.

The above definition controls disparate treatment [41, 3], but it is susceptible to disparate impact [7], which is caused by the proxy features[1] that are highly correlated with S. Therefore, a more vital measure is needed to control indirect discrimination, which is achieved by Demographic Parity (DP) or statistical parity.

Definition 2: Demographic Parity (DP) [3] - Suppose we have a function $f : X \rightarrow \hat{Y}, \hat{Y} = \{0, 1\}$ for binary classification, and let S splits X into a majority set $\mathcal{M}$ and a minority set $\mathcal{M}'$ $(X = \mathcal{M} \cup \mathcal{M}')$, the function f satisfies DP if $P[f(x) = 1 \mid x \in \mathcal{M}] = P[f(x) = 1 \mid x \in \mathcal{M}']$, where x denotes an instance of X and $P[.]$ denotes the probability of an instance. We assume the protected attribute is binary for notational convenience and can be extended to non-binary settings as well.

2.2 Fairness Objective

Most of the state-of-the-art techniques for fairness penalty computations depend on mutual information-related measures [31, 28, 26]. These information-theoretic methods achieve fairness at the expense of data quality and utility. Another line of research is based on adversarial approaches [23, 8] but it suffers from training instability since an adversary cannot be completely trained until convergence in most situations [26]. To tackle these issues, a distance correlation measure has been introduced into the literature [18, 13]. The dependence between two random variables Z_1 and Z_2 can be reduced by minimizing the distance correlation, $\mathcal{V}^2$ between them as follows:

$$\mathcal{V}^2(z_1, z_2) = \int_{\mathcal{Z}_1} \int_{\mathcal{Z}_2} \mid p(z_1, z_2) - p(z_1) p_(z_2) \mid^2 \; dz_1 \; dz_2. \quad (1)$$

2.3 Generative Models

The models in synthetic data generation are mostly based on GANs and DMs. In the proposed work, we use two GAN architectures, namely Least Square GAN [24] and Wasserstein GAN with Gradient Penalty (WGAN-GP) [12] as these are the best among the state-of-the-art GAN-based generation methods. For notational convenience (in this section), let X be the real data and x be an instance of X.

LSGAN. The min-max optimization of LSGAN can be defined as [24]:

$$\begin{aligned} \min_{\theta_D} V_{LSGAN}(D) = -\frac{1}{2} \times \mathbb{E}_{x \sim p_x(X)}\left[(D(x) - b)^2\right] + \\ \frac{1}{2} \times \mathbb{E}_{\xi \sim p_\xi(\xi)}[(D(G(\xi)) - a)^2], \end{aligned} \quad (2)$$

$$\min_{\theta_G} V_{LSGAN}(G) = \frac{1}{2} \times \mathbb{E}_{\xi \sim p_\xi(\xi)}[(D(G(\xi)) - c)^2], \quad (3)$$

where a and b are labels for fake data and real data respectively and c denotes the value that the generator wants D to believe for fake data. Also, ξ is from a uniform or Gaussian distribution $p_\xi(\xi)$ that maps ξ to the real data space through $G(\xi, \theta_G)$.

WGAN-GP. The objective function for WGAN-GP [12] is:

$$L = \underbrace{\underset{\tilde{\boldsymbol{x}} \sim p_g}{\mathbb{E}}[D(\tilde{\boldsymbol{x}})] - \underset{\boldsymbol{x} \sim p_x}{\mathbb{E}}[D(\boldsymbol{x})]}_{\text{Original critic loss}} + \underbrace{\lambda \underset{\hat{\boldsymbol{x}} \sim p_{\hat{\boldsymbol{x}}}}{\mathbb{E}}\left[\left(\|\nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}})\|_2 - 1\right)^2\right]}_{\text{gradient penalty}}, \quad (4)$$

The $p_{\hat{x}}$ denotes sampling uniformly between distribution p_x and the generator distribution p_g. The penalty coefficient $\lambda = 10$ is set according to [2].

Diffusion-based generation. A Diffusion Model (DM) [15] consists of a forward process, in which the data is progressively noised, and a reverse process is applied, in which noise is transformed back into data from the target distribution.

The sampling chain transitions in the forward process can be set to conditional Gaussians and the Markov assumption of the forward process can be defined as [15]:

$$\begin{aligned} \mathbf{q}(X_{1:T} X_0) :=& \prod_{t=1}^{T} \mathbf{q}(X_t X_{t-1}) \\ :=& \prod_{t=1}^{T} \mathcal{N}(X_t; \sqrt{1-\beta_t} X_{t-1}, \beta_t \mathbf{I}), \end{aligned} \quad (5)$$

where $\beta 1, \ldots, \beta T$ is the variance schedule. During the reverse process, the models learn to generate new data starting with the Gaussian noise $\mathbf{p}(X_T) := \mathcal{N}(X_T, \mathbf{0}, \mathbf{I})$, to the joint distribution $\mathbf{p}_\theta(X_{0:T})$ as [15]:

$$\begin{aligned} \mathbf{p}_\theta(X_{0:T}) :=& \mathbf{p}(X_T) \prod_{t=1}^{T} \mathbf{p}_\theta(X_{t-1} X_t) \\ :=& \mathbf{p}(X_T) \prod_{t=1}^{T} \mathcal{N}(X_{t-1}; \boldsymbol{\mu}_\theta(X_t, t), \boldsymbol{\Sigma}_\theta(X_t, t)), \end{aligned} \quad (6)$$

where the time-dependent parameters of the Gaussian transitions are learned.

3 Synthetic Data Fairness

Synthetic data fairness means generating fair synthetic data from biased data so that the downstream models trained on fair synthetic data will have fair predictions in real data [2]. This section forward,

[1] features that are highly correlated with sensitive attributes

[2] We assume that the model does not exhibit explicit biases and the biased outcome is caused only by the biases in the training data.

we separate S from X by $\bar{X} = X \setminus S$ and we write $X \longleftarrow \bar{X}$ for simplicity. Also, we consider $Y \in X$ if not explicitly defined.

Given a biased dataset $\mathbf{D} = \{X, S\}$, where $X \in \mathcal{X}$ and $S \in \mathcal{S}$ respectively denote the set of non-sensitive and sensitive attributes. We define a Fair Data Generation Process (FDGP) as follows:

Definition 3: Fair Data Generation Process (FDGP) - Let $\mathcal{G}$ be a generative model and $\mathcal{U}(S, Y)$ (either FTU or DP) be a definition of algorithmic fairness. The DGP is said to be fair if the $\mathcal{G}$, once optimized, learned to obtain a deterministic transformation from Multivariate Normal Distribution (MVN) to real data distribution that is maximally discriminative with respect to any downstream predictions but invariant to S, evaluated by $\mathcal{U}(S, Y)$.

Definition 4: Synthetic Data Fairness Problem (SDFP) - The Synthetic Data Fairness problem is to generate fair data $\mathbf{D}'$ from biased data $\mathbf{D}$ through a Fair Data Generation Process (FDGP).

In summary, we learn to generate a distribution $p(\mathbf{D}')$ from $p(\mathbf{D})$ by removing the direct and indirect effects of malignant feature S (including proxy attributes), so that the $p(\mathbf{D}')$ can be used for any downstream fair ($\mathcal{U}(S, Y)$ - *fair*) prediction tasks.

4 Fair Latent Deep Generative Models (FLDGMs)

In our proposed FLDGMs, the notion of FDGP is achieved by applying a sequence of operations in the biased data $\mathbf{D}$, which transforms the distribution $p(\mathbf{D})$ to $p(\mathbf{D}')$. This involves separating fairness optimization from data generation while maintaining quality, fairness, and diversity in the target, with the sub-goal of syntax and model-agnostic architectures. The generative models in FLDGMs operate on a comparatively lower dimensional space than that of real data. The framework of FLDGMs can thus be divided into a sequence of three stages:

1. Compressing the real data $\mathbf{D}$ into a fair representation retaining all the necessary information for any target tasks, called fair abstract compression. The output of this stage is low-dimensional fair latent continuous vectors without having any malignant information about sensitive features;
2. Fair latent vector generation, where the generative models are trained to generate high-quality fair latent vectors without focusing on syntax-related information;
3. A high fidelity reconstruction, where the data $\mathbf{D}'$ is reconstructed from the generated fair latent vectors in a single pass.

An outline of our proposed work is given in Figure 1.

4.1 Fair Abstract Compression

Removing undesired variations from data can be considered a general compression model which relies on two sources: a sensitive variable S, which denotes the nuisance we want to remove, and a continuous latent vector Z, which models all the remaining information from input. This fair abstract compression stage can use any of the state-of-the-art fair representation learning methods [42, 8, 29, 18, 23, 13] that consist of an autoencoder trained for the combination of fairness loss and reconstruction loss. The choice of fairness objective in this step depends on different target fairness constraints. This ensures that the fairness optimization has a clear boundary on the attribute space, which is essential as fairness constraints are mostly defined on the independence criteria of attributes.

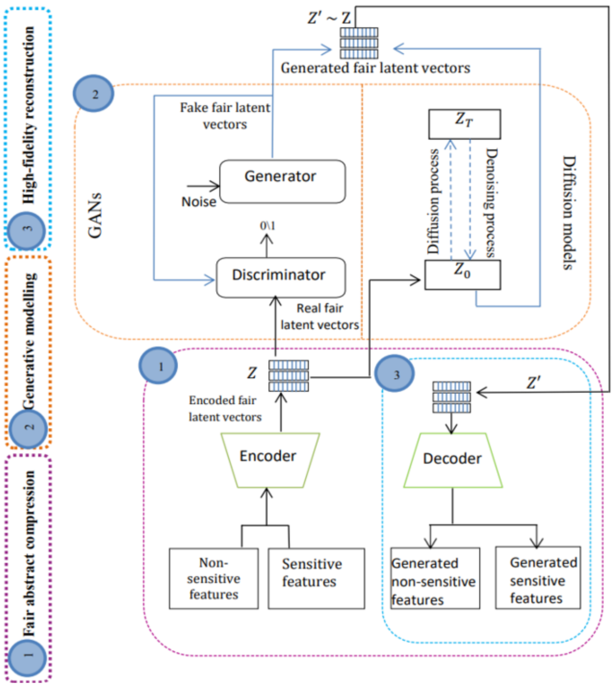


Figure 1: The proposed FLDGM architecture.

Formally, given an input instance $\{(x_i, s_i)\}_{i=1}^{N}$, where N is the number of instances in the data, the encoder $\mathcal{E}$ in the fair abstract compression stage encodes $\{(x_i, s_i)\}_{i=1}^{N}$ into a continuous latent space $z \in$ Z, where $z = \mathcal{E}(x_i, s_i)$. This latent representation Z factors out all the undesired variations in the data about $s \in S \in \mathcal{S}$ using the fairness penalty computed using the distance correlation in Eq. (1) and captures the remaining information for downstream tasks. The decoder $\mathcal{D}$ associated with the autoencoder is now responsible for reconstructing the original data points from the fair latent space, resulting in $\mathcal{D}(z) = \mathcal{D}(\mathcal{E}((x_i, s_i)))$. The fair compression process can be formally defined by :

$$Encode : \mathcal{E}_{\phi_\mathcal{E}}(Z|X, S); Decode : \mathcal{D}_{\phi_\mathcal{D}}(X'|Z, S)), \tag{7}$$

where ϕ is the parameter for the autoencoder and $X \sim X'$. A multivariate Gaussian has been used for posterior $\mathcal{E}_{\phi_\mathcal{E}}(Z|X, S) = \mathcal{N}_{\phi_\mathcal{E}}(Z; \mu, \sigma)$, and a standard multivariate Gaussian $\mathcal{N}(0, I)$ for $p(Z)$.

Therefore, the whole process of the fair abstract compression stage can be defined as a combination of reconstruction loss and fairness loss (from Eq. (1)) [18]:

$$\max_{\phi_\mathcal{E}\phi_\mathcal{D}} \{\log p_{\phi_\mathcal{D}(x|s)} - \alpha \mathcal{V}_\phi^2(z, s)\}, \tag{8}$$

where $\mathcal{V}_\phi^2(z, s)$ is the required independence between the encoded latent space Z and sensitive attributes S and α is a hyperparameter. We analyze the performance of different values of α, $\alpha \in \{1, 2, ..., 10\}$ on fairness and utility and set $\alpha = 7$ for the entire training as it balances the fairness-quality tradeoff as shown in Appendix B.

4.2 Fair Latent Vector Generation

Now, we are attributed with a syntax-agnostic (data type free), fair, and continuous low-dimensional space. Thus, generative frameworks $\mathcal{G}$, such as GANs and DMs, can effectively focus on generating high-quality fair latent vectors without having to deal with high-dimensional feature types (such as categorical features in tabular data and pixels in images). In this work, we use generative models based on GANs and DMs for latent space generation as mentioned in Section

2.3. This can be done by modifying the equations 2-5 by replacing X with our fair latent space Z (and thus by x with z). The only difference is that the real data used to train LSGAN, WGAN-GP, and DM is now the encoded fair latent space Z.

Let Z' be a latent space generated by any of the generative models (LSGAN, WGAN-GP, and DM) then $Z' \sim Z$ for a well-optimized generative model $\mathcal{G}$.

Theorem 1 (Convergence guarantee). *Assume that (i) the data generation is Markov compatible with a pre-trained autoencoder, which is optimized for a combination of fairness loss and reconstruction loss, (ii) the neural networks involved in DGMs have enough capacity, and (iii) the training of all the components of the DGMs is iterative until optimality, then for a well-optimized FLDGM, the generated fair latent distribution $p_{Z'}$ by the generator network $\mathbf{G}$ in $\mathcal{G}$ always converges to the ground-truth fair latent distribution p_Z* (proof in Appendix C).

Theorem 2 (Fairness guarantee). *For a well-optimized generative model $\mathcal{G}$ in FLDGM, the generated fair latent vector Z' is $\mathcal{U}(S, Y)$ - fair, given the corresponding pre-trained autoencoder* (proof in Appendix C).

4.3 High Fidelity Reconstruction

In this stage, we pass the generated fair latent vectors Z' to the pre-trained decoder $\mathcal{D}$ to reconstruct the features from the latent vectors in a single pass. The reconstruction model parameterized by $\phi_{\mathcal{D}}$ can then be represented as:

$$\mathcal{D}_{\phi_{\mathcal{D}}}(X'|Z', S), X' \sim X \tag{9}$$

Summary. The whole process of fair latent deep generative modeling can be formally defined as:

$$\begin{aligned}\mathcal{D}_{\phi_{\mathcal{D}}}&(X'|Z', S) = \mathcal{D}_{\phi_{\mathcal{D}}}(X'|\mathcal{G}_\theta(\xi), Z)\\ &= \underbrace{\mathcal{D}_{\phi_{\mathcal{D}}}(X'|\underbrace{(\mathbf{G}_{\theta_{\mathbf{G}}}(\xi)|\underbrace{(\mathcal{E}_{\phi_{\mathcal{E}}}(Z|X,S)}_{Fair-abstract-compression}))}_{Fair-latent-vector-generation}, S)}_{High-fidelity-reconstruction},\end{aligned} \tag{10}$$

where we denote $\mathcal{G}_\theta$ for any generative model (GAN and DM in our case) and $\mathbf{G}_{\theta_{\mathbf{G}}}$ is the corresponding generator network of $\mathcal{G}_\theta$. Note that, we use ϕ for denoting autoencoder parameters and θ for the generative modeling with an appropriate subscript.

Remark. Given corresponding pre-trained autoencoders, various datasets can be generated based on different fairness constraints and output tasks. This does not add any computational overhead to the generative modeling as fairness is enforced in a separate step.

The term classification fairness [38] or downstream fairness means that any downstream classifiers trained on generated fair data should not discriminate when tested on real data.

Theorem 3 (Classification fairness [38] guarantee). *Any optimal downstream classifiers $\mathbf{M}$ (without any explicit biases) trained on $\mathbf{D}'$ will have fair predictions on $\mathbf{D}$ under $\mathcal{U}(S, Y)$* (proof in Appendix C).

5 Experiments

5.1 Datasets

We performed experiments on three benchmark datasets, Adult Income[3](table), celebA [19], and Color MNIST[17] (image), where the sensitive feature S is significantly correlated with the target label and thus the proper removal of S could be challenging.

Adult Income. The Adult Income is a tabular dataset containing over 65,000 instances with 11 attributes, such as age, education, gender, and income, among others. We treat gender as the sensitive attribute (as there is a known bias between gender and income) and use income as the binary output label representing whether a person earns over $50K$ or not (More details on the analysis of fairness Appendix D).

Color MNIST. The color MNIST is an image database containing handwritten digits and colors for the intrinsic and biased features. Following previous studies, the color MNIST used in our experimental analysis is based on [17]. It is designed with seven standard deviations (SD) (equally spaced between 0.02 and 0.05): the lower the value, the more difficult it is for the model to perform the task since the model can fit the training set by recognizing colors instead of shapes.

CelebA. The CelebA [19] is a large-scale face attributes collection with over 200K celebrity photos with 40 attribute annotations. The photos in this collection span a wide range of pose variants as well as background clutter.

5.2 Evaluation metrics

We evaluate the quality and fairness of our proposed models using the following measures:

1. Data utility - We use precision, recall, and AUROC for evaluating data utility [9, 16, 32]. We train a Random Forest (RF) classifier on synthetic data and test it on real data for downstream prediction and compare it to the state-of-the-art.
2. Sample-level metric - We perform sample-level metrics analysis proposed in [1] to measure the fidelity, diversity, and generalization of synthetic data generated by our proposed models.
3. Fairness - We use both FTU and DP (Section 2.1) for analyzing downstream fairness using a Random Forest classifier.

Furthermore, we performed explainability [22] and bias amplification [35] analysis to substantiate our study. We generated synthetic data using our generative models WGAN-GP, LSGAN, and DM on the datasets mentioned above and computed the metrics by taking an average of over 10 repetitive runs. We have the following variants: FLD-WGAN-GP, FLD-LSGAN, and FLD-DM each with FTU and DP for the fairness definitions in $\mathcal{U}(\mathcal{S}, \mathcal{Y})$. The neural network architectures and implementation details are given in Appendix E.

Competing Methods. The methods we benchmark against for Adult Income data are FairGAN and DECAF (as these models are designed for tabular data). Also, we compare the results of WGAN-GP without fairness to analyze the importance of FDGP. We follow the results from [33][4] as it is difficult to reproduce the results of DECAF as studied in [34]. Additionally, for Color MNIST and CelebA, we performed visualization analysis for verifying the utility, fairness, and quality of our proposed models.

6 Results

6.1 Data Utility and Fairness

6.1.1 Adult Income

We list the utility and fairness measures in Table 1. The precision of our proposed GAN and DM models is far better than FairGAN,

[3] https://archive.ics.uci.edu/ml/datasets/adult

[4] The results are taken from the paper directly

GAN, and WGAN-GP, whereas we obtain almost the same score with DECAF. We improve all the state-of-the-art methods in terms of recall score. The AUROC score of our FLDGM models is better (around 10 percent improvement) compared to the corresponding fair generation methods. Note that DECAF-ND is simply a causal GAN without any fairness optimization. In summary, our FLDGMs achieve a good balance between data quality and fairness compared to state-of-the-art fair generation methods in Adult income data.

6.1.2 Color MNIST

We performed visualization analysis on Color MNIST data. Following [18], we set color as the sensitive attribute, and the generated images are de-correlated from color. Then, we control the color intensity to generate digits with a single color, meaning that the color is disentangled from the digits (this is done in the fair abstract compression stage). By changing the value of color intensity, the generated digits can be either blue, green, or red with approximately the same digit style (Figure 2). Note that, digit generation with fairness does not degrade the image quality, since the generative models in FLDGMs could focus only on image quality in the DGP.

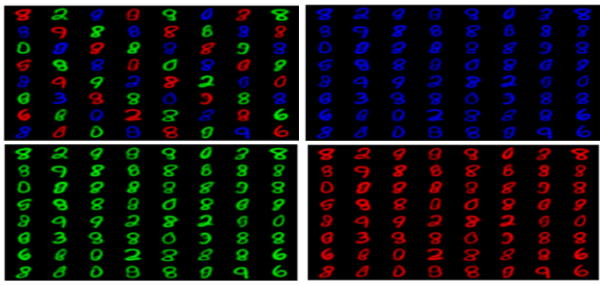
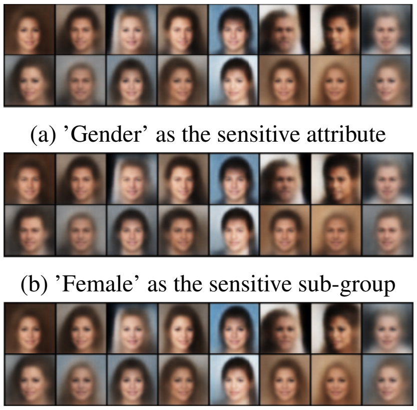


Figure 2: Generated digits with similar styles, color as a sensitive attribute.



(a) ’Gender’ as the sensitive attribute

(b) ’Female’ as the sensitive sub-group

(c) ’Male’ as the sensitive sub-group

Figure 3: PCA modes of generated faces

To verify the quality and fairness of downstream tasks, we train a classifier on the generated digits and tested it on real data, that contains various color biases (by changing the Standard Deviation (SD)) (Table 2.). We get an accuracy of 0.951 for color digit classification, whereas the accuracy of the classifier trained on real data is 0.931. This implies that the proposed FLDGMs helped to improve the classification performance of digit prediction.

6.1.3 CelebA

For the visualization analysis, we focus on the target attribute ’hair color’ and sensitive attribute ’gender’ with “Male” and “female” as sub-groups. We used the same generative model architecture as we used for CMNIST and Adult Income as it operates on the latent vector. We set the training for 10K epochs as the dataset is large and the model converged at 3000 epochs. We consider three combinations { hair color, gender}, { hair color, male}, and { hair color, female} to see how image generation is varied under different sensitive groups and sub-groups. It is interesting to see that when choosing { hair color, gender}, the DP computed between male and female is zero for the prediction. Also, when reducing the correlation between Z and ’male’, all the images generated are females and vice versa and thus it helps to generate desired fairness distributions with appropriate sensitive features (Figure 3).

6.2 Analysis of Density, Coverage, and Accesibility

Additionally, we analyzed the quality of our proposed models in Table 3 using density and coverage metrics [27], and accessibility using the number of parameters used by the models. The reduced parameters mean we need less memory to load these models, increasing accessibility. Note that our proposed models achieve a good balance of density, coverage, and accessibility (measured in no.of parameters) in all three datasets from tabular and image domains. The models based on diffusion achieve better density and coverage as diffusion models are superior in generation compared to GAN-based schemes.

6.3 Sample-level Metric Analysis

Motivated by a recent study [1] on evaluating the faithfulness of synthetic data, we performed sample-level metrics analysis on our proposed variants. This is to measure the quality of synthetic data generation in terms of fidelity, diversity, and authenticity. The results are given in Figure 4. Note that, our proposed models are highly authentic as per [1], which shows the significance of our FLDGMs.

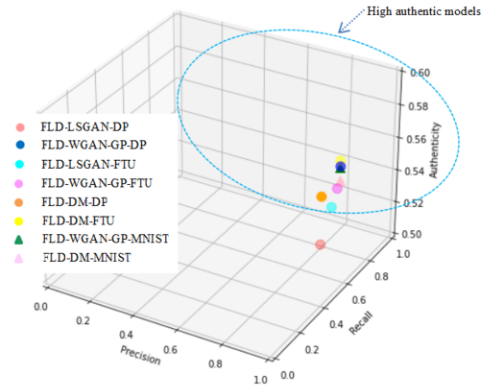


Figure 4: Sample level metrics analysis

6.4 Data Leakage Analysis

To evaluate the bias amplification in terms of data leakage, λ_D and model leakage, λ_M [35] of our proposed model, we train an attacker (which is a ’gender’ classifier on Adult Income data) on the ground truth labels and model predictions by a Random Forest classifier. The

	Data Quality			**Fairness**	
Method	Precision (↑)	Recall (↑)	AUROC (↑)	FTU (↓)	DP (↓)
Real data	0.920 ± 0.006	0.936 ± 0.008	0.807 ± 0.004	0.116 ± 0.028	0.180 ± 0.010
GAN	0.607 ± 0.080	0.439 ± 0.037	0.567 ± 0.132	0.023 ± 0.010	0.089 ± 0.008
WGAN-GP	0.683 ± 0.015	0.914 ± 0.005	**0.798 ± 0.009**	0.120 ± 0.014	0.189 ± 0.024
FairGAN	0.681 ± 0.023	0.814 ± 0.079	0.766 ± 0.029	0.009 ± 0.002	0.097 ± 0.018
DECAF-ND	0.780 ± 0.023	0.920 ± 0.045	0.781 ± 0.007	0.152 ± 0.013	0.198 ± 0.013
DECAF-FTU	0.763 ± 0.033	0.925 ± 0.040	0.765 ± 0.010	0.004 ± 0.004	0.054 ± 0.005
DECAF-CF	0.743 ± 0.022	0.875 ± 0.038	0.769 ± 0.004	0.003 ± 0.006	0.039 ± 0.011
DECAF-DP	0.781 ± 0.018	0.881 ± 0.050	0.672 ± 0.014	0.001 ± 0.002	0.001 ± 0.001
FLD-LSGAN-FTU (ours)	0.762 ± 0.002	**0.998 ± 0.023**	0.762 ± 0.012	0.002 ± 0.001	**0.000 ± 0.001**
FL-LSGAN-DP (ours)	0.763 ± 0.001	0.941 ± 0.002	0.771 ± 0.010	**0.000 ± 0.001**	**0.000 ± 0.000**
FL-WGAN-GP-FTU (ours)	0.772 ± 0.034	0.918 ± 0.001	0.763 ± 0.023	0.001 ± 0.001	**0.000 ± 0.001**
FL-WGAN-GP-DP (ours)	0.782 ± 0.001	0.951 ± 0.001	0.762 ± 0.013	**0.000 ± 0.000**	**0.000 ± 0.000**
FL-DM-FTU (ours)	**0.791 ± 0.011**	0.912 ± 0.002	0.795 ± 0.001	**0.000 ± 0.000**	0.001 ± 0.000
FL-DM-DP (ours)	0.786 ± 0.002	0.905 ± 0.001	0.787 ± 0.011	**0.000 ± 0.001**	**0.000 ± 0.001**

Table 1: Data quality and fairness analysis of the proposed FLDGMs with real data as a reference.

	SD=0.020	SD=0.025	SD=0.030	SD=0.035	SD=0.040	SD=0.045	SD=0.050
($alpha$=0)	.476± .005	.576±.001	.664± .007	0.720±.010	.785± .003	0.838±0.002	.931± .001
(α=0.5)	.901± .001	0.927±.003	.950± .020	0.812±.002	.950±.001	.951±.001	.951±.001

Table 2: Experiment on CMNIST data with FLDGM ($\alpha = 0.5$) and no fairness ($\alpha = 0$)

Datasets	Models	Density (↑)	Coverage (↑)	No.of parameters (↓)
Adult Income	WGAN	0.70937	0.63608	4.756M
	FLD-WGAN-GP (ours)	1.05276	0.80894	**0.224M**
	FLD-DM (ours)	**1.2640 9**	**0.891206**	0.272M
Color MNIST	DCGAN[30]	0.92018	0.70608	5.405M
	FLD-WGAN-GP (ours)	1.01763	0.81290	**0.224M**
	FLD-DM (ours)	**1.19682**	**0.81473**	0.272M
CelebA	DCGAN[30]	0.92318	0.53491	6.342M
	Diffusion	1.29112	0.89196	274M
	FLD-WGAN-GP (ours)	1.10187	0.80537	**0.224M**
	FLD-DM (ours)	**1.29129**	**0.89203**	0.272M

Table 3: Density, coverage, and accessibility analysis of proposed models (↑ - higher the better, ↓ - lower the better, and NA-not available)

λ_{M} trained on different data is given in Table 4. It shows that the leakage is controlled in data generation by all the proposed models as the bias amplification $\Delta(\lambda_{\mathrm{M}} - \lambda_{\mathrm{D}})$ is less than 0 for all the models with both FTU and DP as target fairness constraints.

6.5 Explainability Analysis

In order to analyze the difference in predictions of a Random Forest classifier trained on both real data and synthetic data, generated by our proposed model (we consider FLD-WGAN-GP-DP), we explain the predictions using Shapely Additive Explanation for Adult income data. It is obvious from Figure 5 that the contribution of the feature 'sex' is reduced, whereas the contribution of 'Relationship', 'Educational-Num', and 'Hours per week' (these are intrinsic features for income prediction) is slightly increased.

7 Related Works

We focus on the related literature in terms of (i) non-parametric generative models and (ii) fair synthetic data generation. We refer to Appendix F for an overview of comparing various generative models with respect to our key areas of interest.

7.1 Non-parametric Generative Models

The state-of-the-art methods in synthetic data generation are either based on GANs [12, 40, 37] or Variational Auto Encoders (VAE) [36]. Recently, diffusion models have shown many improvements in high-quality synthetic data generation, particularly in images. The models above are well known for synthetic data generation, having trade-offs in various properties such as quality, diversity, etc, but unable to generate fair data (except [38] as discussed below).

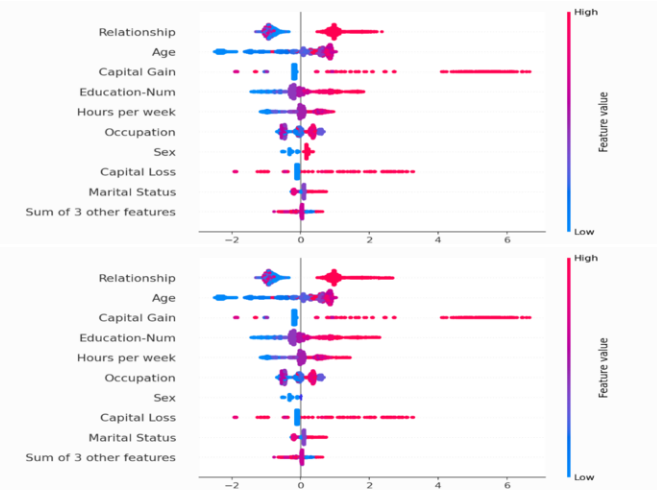


Figure 5: SHAP analysis on Adult Income

Method	λ_M (↓)	Δ(↓)
Real data	0.642	0.022
WGAN-GP	0.712	0.092
FairGAN	0.583	-0.037
FLD-LSGAN-FTU	0.535	-0.085
FLD-LSGAN-DP	0.503	-0.117
FLD-WGAN-GP-FTU	0.505	-0.115
FLD-WGAN-GP-DP	0.502	-0.118
FLD-DM-FTU	**0.500**	**-0.120**
FLD-DM-DP	0.511	-0.109

Table 4: Bias amplification by models trained on different data. The reference dataset leakage is 0.620 with an approximate F1 score of 0.86.

7.2 *Fair Synthetic Data Generation*

The methods under this category [38, 6] range from training generative models with a combination of generative loss and fairness loss, adversarial training, and post-processing schemes.

In FairGAN [38], an adversarial approach is proposed to predict the sensitive attributes from the generated data. This encourages the generator to produce data that is independent of the sensitive features. One main problem with this approach is that the adversary cannot be trained until convergence in every epoch which in turn degrades the performance. Also, this method is designed for binary-sensitive attributes. A fair data generation method by giving access to a small reference fair data is introduced in [6]. The motivation of this work is not aligned with downstream fairness and explicit notions of fairness [33]. A post-processing de-biasing method based on causal knowledge is proposed in [33], where the de-biasing was done at the inference time after the sequential generation of features by individual generators. This approach is designed for tabular data and is strictly based on the causal relationship between features. The computational complexity of this approach is very high, though the de-biasing is flexible to various fairness constraints at the target domain. Another approach based on VAE is proposed in [21], where fairness is introduced by an additional regularization based on Maximum Mean Discrepancy (MMD) to get complete independence between data and sensitive attributes. This method imposes additional overhead for optimizing the MMD in the DGP.

8 Disadvantages and Societal Implications

Disadvantages. The fairness and quality of synthetic data generated by our proposed Fair Latent Deep Generative Models are limited by the performance of the Fair abstract compression stage. Thus the choice of auto-encoder architecture and corresponding fair compression should be designed in a way to balance the tradeoffs between quality and fairness. However, we have succeeded in reducing the computational overhead of syntax-specific generation (high-dimensional) and prevented quality loss when optimizing for fairness in the DGP, with very high flexibility in fine-tuning to various architectures and tasks, which is a great improvement in this context.

Societal Implications. Adversarial attacks on GANs can reveal training instances [5], which is a hot topic of research. However, the extent to which it applies to diffusion models is under-explored. Moreover, generative models tend to exacerbate biases that are present in the training data [14]. In our proposed approach, the training instances are continuous fair latent vectors that do not directly reveal personal information in adversarial attacks (as it is encoded). Therefore, in an environment where privacy is of great concern, it is advisable to have an authentic human-in-the-loop who keeps the details of the autoencoder and shares other components for downstream applications, thereby having proper control over data privacy. For the second problem, de-biasing data happens in the fair compression stage, thereby the subsequent generative modeling could not access the bias information in the data, which greatly controls the bias amplification in downstream models.

9 Conclusion

We have proposed Fair Latent Deep Generative Models (FLDGMs), a syntax-agnostic generative framework that enables an efficient way to significantly improve both the fairness and quality of synthetic data generation using Diffusion models and Generative Adversarial Networks on image and tabular data. Based on our experimental analysis and evaluation, we demonstrated favorable results in terms of data quality, authenticity, accessibility, and fairness compared to state-of-the-art schemes across a wide range of proposed models in the absence of task-specific architectures.

Future Directions. One interesting future research direction could be to extend this framework for de-biasing hate speech detection and replace the biased contents in the tweets or speech with another, that could be generated by any underlying Natural Language Generation (NLG) methods. This area has not been explored but is very important as it has applications in text summarization, question generation, hate speech detection and removal, and text-to-image generation. Also, in contexts, where multi-modal data contains various biases, it could be interesting to first learn a common representation without biases and then build downstream models on top of it.

Acknowledgments

This work was funded by the Knut and Alice Wallenberg Foundation, the ELLIIT Excellence Center at Linköping-Lund for Information Technology, and TAILOR - an EU project with the aim to provide the scientific foundations for Trustworthy AI in Europe. The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre. The authors would also like to thank Dr. Daniel de Leng, Department of Computer and Information Science, LiU for the insightful comments.

References

[1] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar, 'How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models', in *International Conference on Machine Learning*, pp. 290–306. PMLR, (2022).

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou, 'Wasserstein generative adversarial networks', in *International conference on machine learning*, pp. 214–223. PMLR, (2017).

[3] Solon Barocas and Andrew D Selbst, 'Big data's disparate impact', *Calif. L. Rev.*, **104**, 671, (2016).

[4] Reuben Binns, 'Fairness in machine learning: Lessons from political philosophy', in *Conference on Fairness, Accountability and Transparency*, pp. 149–159. PMLR, (2018).

[5] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al., 'Extracting training data from large language models', in *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, (2021).

[6] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon, 'Fair generative modeling via weak supervision', in *International Conference on Machine Learning*, pp. 1887–1898. PMLR, (2020).

[7] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian, 'Certifying and removing disparate impact', in *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, (2015).

[8] Rui Feng, Yang Yang, Yuehan Lyu, Chenhao Tan, Yizhou Sun, and Chunping Wang, 'Learning fair representations via an adversarial framework', *arXiv preprint arXiv:1904.13341*, (2019).

[9] Peter Flach and Meelis Kull, 'Precision-recall-gain curves: Pr analysis done right', *Advances in neural information processing systems*, **28**, (2015).

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, 'Generative adversarial networks', *Communications of the ACM*, **63**(11), 139–144, (2020).

[11] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller, 'The case for process fairness in learning: Feature selection for fair decision making', in *NIPS symposium on machine learning and the law*, volume 1, p. 2. Barcelona, Spain, (2016).

[12] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, 'Improved training of wasserstein gans', *Advances in neural information processing systems*, **30**, (2017).

[13] Dandan Guo, Chaojie Wang, Baoxiang Wang, and Hongyuan Zha, 'Learning fair representations via distance correlation minimization', *IEEE Transactions on Neural Networks and Learning Systems*, (2022).

[14] Aman Gupta, Deepak Bhatt, and Anubha Pandey, 'Transitioning from real to synthetic data: Quantifying the bias in model', *arXiv preprint arXiv:2105.04144*, (2021).

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel, 'Denoising diffusion probabilistic models', *Advances in Neural Information Processing Systems*, **33**, 6840–6851, (2020).

[16] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila, 'Improved precision and recall metric for assessing generative models', *Advances in Neural Information Processing Systems*, **32**, (2019).

[17] Jungsoo Lee, Eungyeup Kim, Juyoung Lee, Jihyeon Lee, and Jaegul Choo, 'Learning debiased representation via disentangled feature augmentation', *Advances in Neural Information Processing Systems*, **34**, 25123–25133, (2021).

[18] Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong, 'Fair representation learning: An alternative to mutual information', in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1088–1097, (2022).

[19] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, 'Deep learning face attributes in the wild', in *Proceedings of International Conference on Computer Vision (ICCV)*, (December 2015).

[20] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel, 'The variational fair autoencoder', *arXiv preprint arXiv:1511.00830*, (2015).

[21] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S Zemel, 'The variational fair autoencoder', in *ICLR*, (2016).

[22] Scott M Lundberg and Su-In Lee, 'A unified approach to interpreting model predictions', *Advances in neural information processing systems*, **30**, (2017).

[23] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel, 'Learning adversarially fair and transferable representations', in *International Conference on Machine Learning*, pp. 3384–3393. PMLR, (2018).

[24] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley, 'Least squares generative adversarial networks', in *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, (2017).

[25] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan, 'A survey on bias and fairness in machine learning', *ACM Computing Surveys (CSUR)*, **54**(6), 1–35, (2021).

[26] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg, 'Invariant representations without adversarial training', *Advances in Neural Information Processing Systems*, **31**, (2018).

[27] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo, 'Reliable fidelity and diversity metrics for generative models', in *International Conference on Machine Learning*, pp. 7176–7185. PMLR, (2020).

[28] Lihao Nan and Dacheng Tao, 'Variational approach for privacy funnel optimization on continuous data', *Journal of Parallel and Distributed Computing*, **137**, 17–25, (2020).

[29] Luca Oneto, Michele Donini, Massimiliano Pontil, and Andreas Maurer, 'Learning fair and transferable representations with theoretical guarantees', in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 30–39. IEEE, (2020).

[30] Alec Radford, Luke Metz, and Soumith Chintala, 'Unsupervised representation learning with deep convolutional generative adversarial networks', *arXiv preprint arXiv:1511.06434*, (2015).

[31] Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund, 'A variational approach to privacy and fairness', in *2021 IEEE Information Theory Workshop (ITW)*, pp. 1–6. IEEE, (2021).

[32] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly, 'Assessing generative models via precision and recall', *Advances in neural information processing systems*, **31**, (2018).

[33] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar, 'Decaf: Generating fair synthetic data using causally-aware generative networks', *Advances in Neural Information Processing Systems*, **34**, 22221–22233, (2021).

[34] Shuai Wang, Paul Verhagen, Jennifer Zhuge, and Velizar Shulev, 'Replication study of decaf: Generating fair synthetic data using causally-aware generative networks', in *ML Reproducibility Challenge 2021 (Fall Edition)*, (2022).

[35] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez, 'Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations', in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5310–5319, (2019).

[36] Max Welling and Diederik P Kingma, 'Auto-encoding variational bayes', *ICLR*, (2014).

[37] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou, 'Differentially private generative adversarial network', *arXiv preprint arXiv:1802.06739*, (2018).

[38] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu, 'Fairgan: Fairness-aware generative adversarial networks', in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE, (2018).

[39] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu, 'Fairgan+: Achieving fair data generation and classification through generative adversarial nets', in *2019 IEEE International Conference on Big Data (Big Data)*, pp. 1401–1406. IEEE, (2019).

[40] Jinsung Yoon, Lydia N Drumright, and Mihaela Van Der Schaar, 'Anonymization through data synthesis using generative adversarial networks (ads-gan)', *IEEE journal of biomedical and health informatics*, **24**(8), 2378–2388, (2020).

[41] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi, 'Fairness constraints: Mechanisms for fair classification', in *Artificial intelligence and statistics*, pp. 962–970. PMLR, (2017).

[42] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork, 'Learning fair representations', in *International conference on machine learning*, pp. 325–333. PMLR, (2013).