Fair-CDA: Continuous and Directional Augmentation for Group Fairness

Rui Sun[1,2*], Fengwei Zhou[3*], Zhenhua Dong[3], Chuanlong Xie[4†], Lanqing Hong[3], Jiawei Li[3], Rui Zhang[5], Zhen Li[1,2†], Zhenguo Li[3]

[1] The Future Network of Intelligence Institute, The Chinese University of Hong Kong (Shenzhen)
[2] School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen)
[3] Huawei Noah's Ark Lab
[4] Beijing Normal University
[5] Tsinghua University
ruisun@link.cuhk.edu.cn, fzhou@connect.ust.hk, {dongzhenhua, honglanqing, li.jiawei, li.zhenguo}@huawei.com, rayteam@yeah.net, clxie@bnu.edu.cn, lizhen@cuhk.edu.cn

Abstract

In this work, we propose *Fair-CDA*, a fine-grained data augmentation strategy for imposing fairness constraints. We use a feature disentanglement method to extract the features highly related to the sensitive attributes. Then we show that group fairness can be achieved by regularizing the models on transition paths of sensitive features between groups. By adjusting the perturbation strength in the direction of the paths, our proposed augmentation is controllable and auditable. To alleviate the accuracy degradation caused by fairness constraints, we further introduce a calibrated model to impute labels for the augmented data. Our proposed method does not assume any data generative model and ensures good generalization for both accuracy and fairness. Experimental results show that Fair-CDA consistently outperforms state-of-the-art methods on widely-used benchmarks, e.g., Adult, CelebA and MovieLens. Especially, Fair-CDA obtains an 86.3% relative improvement for fairness while maintaining the accuracy on the Adult dataset. Moreover, we evaluate Fair-CDA in an online recommendation system to demonstrate the effectiveness of our method in terms of accuracy and fairness.

Introduction

Many machine learning systems have achieved empirically success in practical problems but may sometimes raise issues of discrimination and unfairness. In job candidate search, different protected groups (e.g., gender and ethnic groups) may be treated unfairly in terms of their members appearing in recommended candidate lists (Ekstrand et al. 2021). In the context of information retrieval, unfairness may happen among multiple parties. For example, unfair exposure allocation may favour monopolies and drive small content providers out of the market (Morik et al. 2020). This reduces diversity and impairs the whole ecosystem.

There have been various studies to impose fairness constraints during training procedure (Zemel et al. 2013; Hardt, Price, and Srebro 2016; Zafar et al. 2017; Chuang and Mroueh 2021), ensuring that different groups shall be treated similarly. However, these constraints are data-dependent, the learnt fair classifiers might not generalize at evaluation time. Agarwal et al. (2018) and Cotter et al. (2019) consider two-player games to formulate the constrained optimization problem and analyze the solutions and generalization guarantees. Chuang and Mroueh (2021) proposes *Fair Mixup* to generate a path of distributions that connects sensitive groups and regularize the smoothness of transitions among the path to improve the generalization of group fairness metrics. They show that their strategy ensures a better generalization for both accuracy and fairness in a wide range of benchmarks.

Motivated by *Fair Mixup*, we propose *Fair-CDA*, a continuous and directional augmentation method, to seek a fine-grained balance between fairness and accuracy. An overview of our method is illustrated in Figure 1.

Accuracy. The *Mixup* (Zhang et al. 2018) generates augmented samples via convex combinations of pairs of data points. However, the between-group augmentation performs *Mixup* on both sensitive attributes and non-sensitive attributes. This may change the correlation between non-sensitive attributes and the targets and further lead to the fall of prediction accuracy. What's more, Verma et al. (2019) shows that interpolations in deeper hidden layers, which capture higher-level information (Zeiler and Fergus 2014), can provide additional training signal and smooth decision boundaries that benefit generalization. Therefore we develop a fine-grained augmentation via feature disentanglement and focus on the transitions over sensitive features.

We decompose representations in latent space into sensitive and non-sensitive features via *DecAug* (Bai et al. 2020), which is a powerful feature disentanglement technique for Out-of-Distribution (OoD) generalization. The sensitive features encode the information that is strongly correlated to the sensitive attributes, while the non-sensitive features retain as much other information (essential for prediction) as possible. Then we apply semantic augmentation on the sensitive features aiming to generate features correlated to the opposite sensitive attributes. The augmented sensitive fea-

*These authors contributed equally.
†Correspondence to: lizhen@cuhk.edu.cn,clxie@bnu.edu.cn

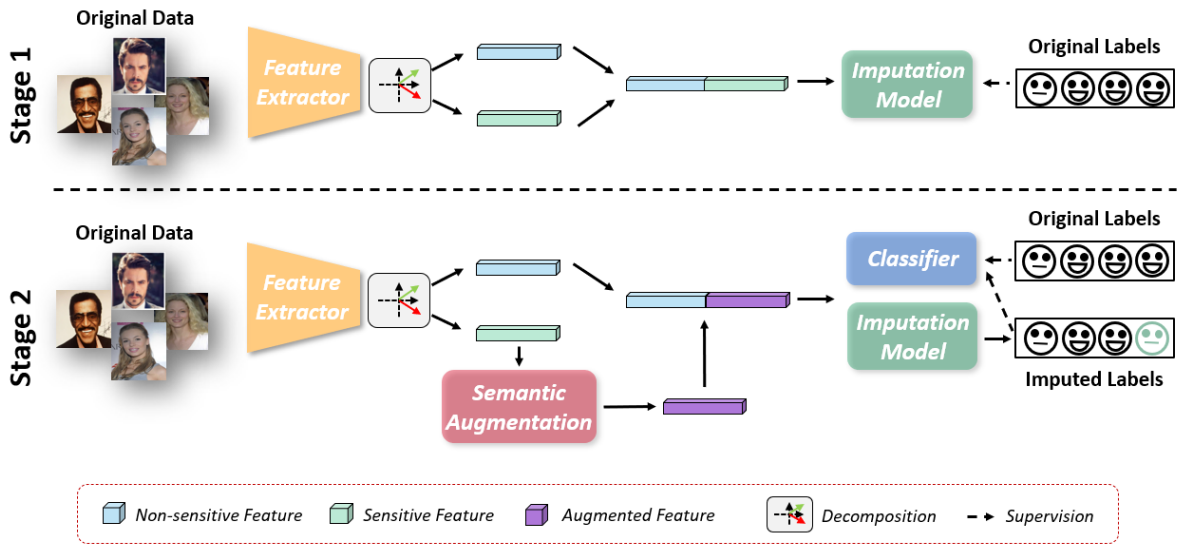


Figure 1: An overview of the proposed Fair-CDA.

tures combined with the original non-sensitive features form what we refer to as the augmented features.

Fairness. We eliminate the disparities of the predictions made by a task model via training the model to make the same decisions for the samples with the original features and with the corresponding augmented features. So the augmentation strategy determines the level of fairness. However, it is not easy to control *Mixup* by tuning the distribution of the interpolation weight, which is usually a Beta distribution. According to Zhang et al. (2021), we consider adversarial training with random perturbation size to augment sensitive features more related to the opposite sensitive attribute. The perturbation budget of the adversary becomes a key hyperparameter that monitors the generation procedure and controls the degree of fairness. When the perturbation is significant, the augmented features can be classified into the opposite group with a high probability. Then the classifier learnt with large perturbations becomes fairer against the sensitive attribute.

A potential competitive edge of our approach is that we can audit a learnt model at the individual level based on objective criteria: whether the task model is robust to the perturbation of the attribute classifier. Given an individual, the key problem for testing discrimination is to simulate the corresponding individual with a different protected attribute. The white-box framework, *Counterfactual Fairness* (Kusner et al. 2017) can accurately detect the bias and understand how the model discriminates. But the reliance on a known causal mechanism limits its application scenarios and may fail to identify instances of legally actionable discrimination. Some black-box techniques generate mirror individuals via a learnt generative model, e.g. *FlipTest* (Black, Yeom, and Fredrikson 2020). Notice that central to assessing fair data generation is a learnt predictor for the sensitive attribute. However, an intuitive question arises: had an individual been of an opposite sensitive attribute, would the attribute predictor output the opposite attribute? Without a white-box assumption, the optimal Bayesian classifier may fail to predict the opposite attribute for some individuals, and then the audits run the risk of becoming circular verification. The proposed method does not suffer from this problem due to the controllable augmentation. The perturbation budget during training presents a quantitative standard to perform the model audit.

We summarize the contributions as follows:

- We propose Fair-CDA that precisely applies augmentation for sensitive features to achieve fairness while compromising little accuracy.
- The proposed augmentation is controllable by tuning the perturbation budget of the adversary and provides an audit criterion via adversarial robustness.
- Extensive experiments show that Fair-CDA significantly outperforms state-of-the-art methods in various settings, which is effective and scalable. For instance, Fair-CDA can be applied to different backbones, different tasks, and real application scenarios.

Preliminaries

Suppose that data points $\{(x_i, y_i)\}$ are drawn according to some unknown joint distribution over $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} \subseteq \mathbb{R}^d$. Let A denote the given sensitive attribute that should not be treated differently in decision-making. Without loss of generality, we consider a binary classification task $\mathcal{Y} = \{0, 1\}$ with a binary sensitive attribute $A \in \{0, 1\}$. Let $\hat{Y}$ be the predictor, a random variable that is produced by a model $f : \mathcal{X} \to [0, 1]$ as a prediction of Y. In this work, we focus on two widely-used group fairness constraints: Demographic Parity (DP) (Dwork et al. 2012) and Equalized Odds (EO) (Hardt, Price, and Srebro 2016).

Demographic Parity. A predictor $\hat{Y}$ satisfies demographic parity (DP) if $P(\hat{Y} \mid A = 0) = P(\hat{Y} \mid A = 1)$. DP requires $\hat{Y}$ to be statistically independent of A. In real-world applications, we use DP as the definition to ensure fairness when historically biased decisions may have affected the quality of the collected data and we want to see minority groups receiving positive decisions at the same rate. To evaluate the fairness of a trained model f under this definition, we use the following relaxed metric (Madras et al. 2018; Chuang and Mroueh 2021):

$$\Delta_{DP}(f) = \big|\mathbb{E}[f(X)|A = 0] - \mathbb{E}[f(X)|A = 1]\big|.$$

To fulfill the requirement of DP, the evaluation metric $\Delta_{DP}(f)$ shall go to zero. Since for certain predictions, such as hobbies or expertise, there are indeed differences between groups, meeting the mandatory requirements of DP will greatly decrease the prediction accuracy. Hence, the following alternate criterion is proposed to overcome the limitations of DP.

Equalized Odds. A predictor $\hat{Y}$ satisfies equalized odds (EO) if $P(\hat{Y} \mid A = 0, Y = y) = P(\hat{Y} \mid A = 1, Y = y)$, for any $y \in \{0, 1\}$. EO requires $\hat{Y}$ to be independent of A conditioned on Y. In real-world applications, we use EO as a criterion to ensure fairness if there are strict requirements for making correct predictions and we strongly care about the qualifications of candidates when making decisions. Similarly, to evaluate the fairness of a trained model f under EO, we use the following metric (Madras et al. 2018; Chuang and Mroueh 2021):

$$\begin{aligned}\Delta_{EO}(f) \quad = \quad & \sum_{y\in\{0,1\}} \Big|\mathbb{E}[f(X)|A = 0, Y = y] \\ & -\mathbb{E}[f(X)|A = 1, Y = y]\Big|.\end{aligned}$$

Different from DP, EO considers the possible correlation between Y and A, it does not rule out the perfect predictor even when the base rates differ across groups.

Fair-CDA

To fulfill the fairness constraint, we shall reduce the dependence of model predictions on sensitive attributes. Simply removing the sensitive attributes from the inputs does not necessarily lead to a non-discriminatory model prediction, as other attributes in the inputs might encode information for inferring the sensitive attributes (Dwork et al. 2012; Feldman et al. 2015). Hence, we need to decompose the representations of the inputs into sensitive and non-sensitive features. Sensitive features encode information that can identify whether the inputs belong to a certain group determined by the sensitive attributes, while non-sensitive features retain as much other information as possible (Zemel et al. 2013). Moreover, we shall obfuscate the sensitive features to obtain a fair model. To decompose the high-level representations of the inputs, we train a task model to predict both data labels and sensitive attributes with an orthogonality constraint on gradients for the intermediate features (Bai et al. 2020).

Feature Disentanglement. Consider a task with training data $\{(x_i, y_i, a_i)\}_{i=1}^n$. To decompose the representations, we denote three feature extractor: h, h_y and h_a, and write their output features as

$$z_i = h(x_i), \quad z_i^y = h_y(z_i), \quad z_i^a = h_a(z_i).$$

The mapping h is a pre-extractor that learns the high-level representations of the input. Then h_a and h_y are two additional extractors after h to obtain sensitive and non-sensitive features. The principle here is to enforce h_y to extract features that affect the label prediction loss the most will not affect the sensitive attribute prediction loss and vice versa. Therefore we design a regularization term as follows:

$$\beta(\mathcal{L}_i^y + \mathcal{L}_i^a + \mathcal{L}_i^\perp), \tag{1}$$

where β is a tuning parameter and

$$\begin{aligned}\mathcal{L}_i^y &:= \mathcal{L}_i^y(h, h_y, g_y) = \ell(g_y(z_i^y), y_i), \\ \mathcal{L}_i^a &:= \mathcal{L}_i^a(h, h_a, g_a) = \ell(g_a(z_i^a), a_i),\end{aligned}$$

and

$$\mathcal{L}_i^\perp := \mathcal{L}_i^\perp(h_y, h_a, g_y, g_a) = \frac{\langle \nabla_{z_i}\mathcal{L}_i^y, \nabla_{z_i}\mathcal{L}_i^a\rangle^2}{\|\nabla_{z_i}\mathcal{L}_i^y\|^2 \cdot \|\nabla_{z_i}\mathcal{L}_i^a\|^2}.$$

Here g_y and g_a are two classifier to predict y and a and ℓ is the cross-entropy loss. The term $\mathcal{L}_i^\perp$ imposes a constraint on gradient orthogonality to disentangle features. To estimate the feature extractors and the classifiers, Stage 1 of Fair-CDA (Figure. 1) minimizes the objective function:

$$\frac{1}{n}\sum_{i=1}^n \mathcal{L}_i + \beta(\mathcal{L}_i^y + \mathcal{L}_i^a + \mathcal{L}_i^\perp),$$

where $\mathcal{L}_i$ is the loss function of the task model g over the disentangled features:

$$\mathcal{L}_i := \mathcal{L}_i(h, h_y, h_a, g) = \ell(g([z_i^y, z_i^a]), y_i). \tag{2}$$

Semantic Augmentation. In Stage 2 of Fair-CDA, we do an intervention on the sensitive features to mitigate unfair biases. Intuitively, a model satisfies the requirement of the fairness constraint if it can make the same prediction for two samples with different sensitive features but the same other features (Kusner et al. 2017). We augment the sensitive features along the direction of increasing the attribute prediction loss $\mathcal{L}_i^a$:

$$\tilde{z}_i^a = z_i^a + \alpha_i \frac{\nabla_{z_i^a}\ell(g_a(z_i^a), a_i)}{\big\|\nabla_{z_i^a}\ell(g_a(z_i^a), a_i)\big\|}, \tag{3}$$

where α_i is a perturbation size. Since the direction of the gradient is the direction in which the loss increases most rapidly, augmentation in this way changes z_i^a to the features corresponding to the other sensitive attribute.

Transition Path. Chuang and Mroueh (2021) interpolates the transition path between groups via Mixup. Zhang et al. (2021) proves that the adversarial loss can be bounded above by the Mixup loss. Therefore we generate the transition path over sensitive features by randomizing the perturbation size. In this work, we assume α_i is a random variable follows a uniform distribution over $[0, \lambda]$. Here λ is the perturbation budget of (3) that controls the strength of the augmentation.

After obtaining the generated sensitive features $\tilde{z}_i^a$, we concatenate them with the non-sensitive features z_i^y and train the task model g with $\{([z_i^y, \tilde{z}_i^a], y_i)\}$. The loss function of g over the augmented features is denoted by

$$\tilde{\mathcal{L}}_i := \tilde{\mathcal{L}}_i(h, h_y, h_a, g) = \ell(g([z_i^y, \tilde{z}_i^a]), y_i). \tag{4}$$

together with the aforementioned losses $\mathcal{L}_i^1$, $\mathcal{L}_i^2$ and $\mathcal{L}_i^\perp$.

Different from existing works (Lahoti, Gummadi, and Weikum 2019; Zafar et al. 2017; Chuang and Mroueh 2021), our method strikes a balance between accuracy and fairness via adjusting the perturbation budget λ. When the perturbation is large, the augmented features can be classified into the opposite group with a high probability. Then the classifier learnt with large perturbations becomes fairer against the sensitive attribute.

Imputation Model. In Stage 2, we use the labels of the original samples to mark the corresponding augmented features. This is based on the intuition that a model is non-discriminatory if it makes the same prediction for two samples only differing in sensitive features. However, for certain predictions, there are indeed correlations between labels and sensitive features. Enforcing the model to meet the mandatory requirement of the fairness constraint and ignoring the possible correlations between labels and sensitive features may decrease the prediction accuracy a lot. To further improve the prediction accuracy, we introduce an imputation model to calibrate the labels of the augmented features. Specifically, the Stage 1 solution of the task model, denoted by $\breve{g}$, is taken to be the imputation model to label the augmented features: $\breve{y}_i = \breve{g}([z_i^y, \tilde{z}_i^a])$. The loss of predicting $\breve{y}_i$ is denoted by

$$\breve{\mathcal{L}}_i = \breve{\mathcal{L}}_i(h, h_y, h_a, g) = \ell(g([z_i^y, \tilde{z}_i^a]), \breve{y}_i). \tag{5}$$

The task model is then trained to predict both the original labels and the labels given by the imputation model. We formulate the final problem of Fair-CDA as minimizing:

$$\frac{1}{n}\sum_{i=1}^{n} \gamma\tilde{\mathcal{L}}_i + (1-\gamma)\breve{\mathcal{L}}_i + \beta(\mathcal{L}_i^1 + \mathcal{L}_i^2 + \mathcal{L}_i^\perp), \tag{6}$$

where γ is a hyper-parameter balancing $\tilde{\mathcal{L}}_i$ and $\breve{\mathcal{L}}_i$. For time-saving, our method initializes with the imputation model to solve the optimization problem in (6).

Summary. As mentioned above, Fair-CDA balances the prediction accuracy and fairness via adjusting the perturbation strength λ. The algorithm is summarized in Algorithm 1. Stage 1 disentangles features and learns the task model with the original training samples. In Stage 2, we fine-tune the task model with the augmented features to achieve fairness.

Our method introduces three additional hyper-parameters: two weights of different losses β and γ, and perturbation budget λ. In the experiment, we set β according to the initial loss values to make different loss values in the same magnitude range. We adjust γ on the Adult dataset (Dua and Graff 2017) to get the best accuracy and fairness trade-off on the validation set and then adopt the same value which is 0.9 for all the datasets. Our method balances the prediction accuracy and fairness via adjusting the perturbation strength λ, while previous works (Chuang and Mroueh 2021; Zhang, Lemoine, and Mitchell 2018) balance them via adjusting the weights of the regularization terms. On different datasets, we first conduct experiments with $\lambda = 0, 1, 10, 100, 1000$ to narrow down the range of λ and then, do grid search between the determined range of λ (reported in Appendix) with a budget of 20 points to generate the Pareto Front in Figure 2,5,6,7. In real-world applications, the number of grid search points can be determined according to the budget.

Algorithm 1: Fair-CDA: Continuous and Directional Augmentation for Group Fairness

Input: Training data $\{(x_i, y_i, a_i)\}_{i=1}^n$, batch sizes b, learning rate η_1, η_2, perturbation strength λ, weights γ, β, iteration number T, S

Output: $\theta = (h, h_y, h_a, g, g_y, g_a)$;

Stage 1:

Initialize $\theta^{(0)} = (h^{(0)}, h_y^{(0)}, h_a^{(0)}, g^{(0)}, g_y^{(0)}, g_a^{(0)})$;

for $1 \le t \le T$ **do**

  Sample a batch of training data $\{(x_i, y_i, a_i)\}_{i=1}^b$;

  Compute $\mathcal{L}_i$, $\mathcal{L}_i^y$, $\mathcal{L}_i^a$, and $\mathcal{L}_i^\perp$ according to Eq. (1) and Eq. (2)

  Update θ

$$\theta^{(t)} = \theta^{(t-1)} - \frac{\eta_1}{b}\sum_{i=1}^{b} \nabla_\theta\big(\mathcal{L}_i + \beta(\mathcal{L}_i^y + \mathcal{L}_i^a + \mathcal{L}_i^\perp)\big);$$

end for

Stage 2:

for $1 \le s \le S$ **do**

  Sample a batch of training data $\{(x_i, y_i, a_i)\}_{i=1}^b$;

  for each (x_i, y_i, a_i) **do**

    Compute $z_i^y = h_y^{(T+s-1)} \circ h^{(T+s-1)}(x_i)$ and $z_i^a = h_a^{(T+s-1)} \circ h^{(T+s-1)}(x_i)$;

    Compute $\mathcal{L}_i^y$, $\mathcal{L}_i^a$ and $\mathcal{L}_i^\perp$;

    Randomly draw α_i according to $U(0, \lambda)$;

    Generate $\tilde{z}_i^a$ according to Eq. (3);

    Compute $\tilde{\mathcal{L}}_i$ according to Eq. (4);

    Impute the label $\tilde{y}_i = g^{(T)}([z_i^y, \tilde{z}_i^a])$;

    Compute $\breve{\mathcal{L}}_i$ according to Eq. (5);

  end for

  Update θ:

$$\theta^{(T+s)} = \theta^{(T+s-1)} - \frac{\eta_2}{b}\sum_{i=1}^{b} \nabla_\theta\big(\gamma\tilde{\mathcal{L}}_i + (1-\gamma)\breve{\mathcal{L}}_i + \beta(\mathcal{L}_i^y + \mathcal{L}_i^a + \mathcal{L}_i^\perp)\big);$$

end for

Experiments on Public Datasets

We evaluate Fair-CDA on tabular dataset Adult (Dua and Graff 2017), vision dataset CelebA (Liu et al. 2018), and recommender dataset MovieLens (Harper and Konstan 2015). We demonstrate the effectiveness of Fair-CDA across diverse tasks and task models. In the ablation studies, we examine the contributions of feature decomposition and the imputation model. We compare Fair-CDA with other baseline methods: ERM, GapReg (Chuang and Mroueh 2021), AdvDebias (Zhang, Lemoine, and Mitchell 2018), and Mixup / Manifold Mixup (Chuang and Mroueh 2021)

Task	Attribute	Label	Ratio
Adult	Female	Salary$<= 50k$	88.7%
	Female	Salary$> 50k$	11.3%
	Male	Salary$<= 50k$	68.5%
	Male	Salary$> 50k$	31.5%
CelebA (Smiling)	Female	Not Smiling	46.2%
	Female	Smiling	53.8%
	Male	Not Smiling	60.1%
	Male	Smiling	39.9%
CelebA (Wavy Hair)	Female	Not Wavy Hair	55.3%
	Female	Wavy Hair	44.7%
	Male	Not Wavy Hair	85.7%
	Male	Wavy Hair	14.3%
CelebA (Attractive)	Female	Not Attractive	31.7%
	Female	Attractive	68.3%
	Male	Not Attractive	72.1%
	Male	Attractive	27.9%
MovieLens	Minority	Not Recommend	41.4%
	Minority	Recommend	58.6%
	Majority	Not Recommend	44.1%
	Majority	Recommend	55.9%

Table 1: Statistical data of different tasks on three datasets.

using two metrics: prediction accuracy and fairness. To measure the accuracy, we use Average Precision (AP) for tabular (Adult) and vision (CelebA) tasks, and Area Under Curve (AUC) for recommender (MovieLens) task. To measure fairness, we use two widely-used fairness metrics: Demographic Parity (DP) and Equalized Odds (EO) which are defined in Preliminaries. Also, we compare our method with FFAVE and β-VAE following the setting in (Creager et al. 2019). Please refer to the Appendix for more details about the datasets.

Unjustified biases from the observed data. We count the imbalance in the number of training data across sensitive attribute groups and the detailed statistical data are shown in Table 1. In the Adult dataset, the proportion of males with high salaries is significantly higher than that of females. In the CelebA dataset, the proportion of males with a positive label is significantly lower than that of females. In the MovieLens dataset, movies from minority producers also have different positive rates from that of another group. All these imbalances and biases can be inherited and amplified by the models.

Implementations. Our framework is implemented with PyTorch 1.4 (under BSD license), Python 3.7, and CUDA v9.0. For the baseline methods, we implement with PyTorch 1.3.1 to keep the same setting as their source code. We conducted experiments on NVIDIA Tesla V100. The results of baseline methods on Adult and CelebA are referenced from (Chuang and Mroueh 2021), while the results of baseline methods on MovieLens are implemented by ourselves.

Results and Discussion

Results on Adult dataset. Fair-CDA achieves State-Of-The-Art (SOTA) performance in terms of both fairness and accuracy on the Adult dataset, as shown in Figure 2. ERM has a moderate AP but poor fairness, while GapReg achieves better fairness but lower AP than ERM. It utilizes the fairness constraint in the training phase, which lacks generalization at evaluation time. Fair Mixup achieves a better trade-off compared to the previous three methods but is dominated by Fair-CDA. In particular, Fair-CDA is the only method consistently achieving a higher AP than ERM under two fairness constraints.

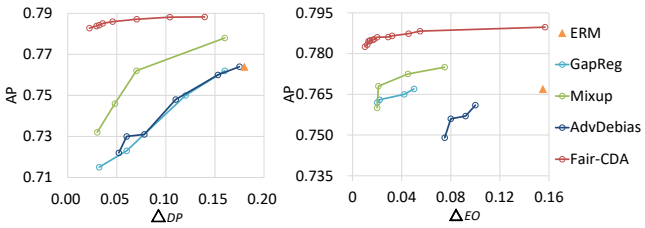


Figure 2: The trade-off between AP and Δ_{DP} / Δ_{EO} on Adult dataset.

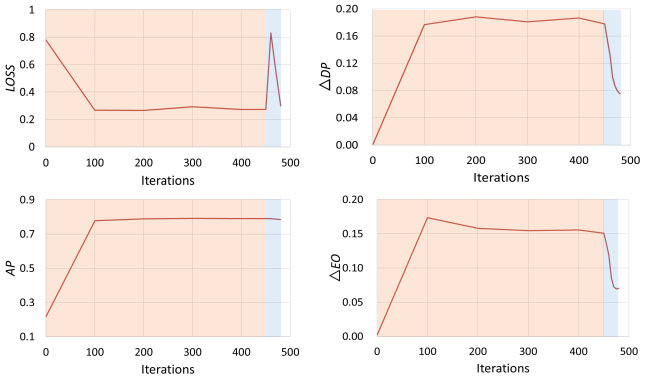


Figure 3: The trend of loss and performance during two stages on Adult dataset. The orange color in the figure corresponds to Stage 1, while the blue color represents Stage 2. In Stage 1, Fair-CDA mainly optimizes the prediction accuracy, while in Stage 2, Fair-CDA mainly optimizes the model fairness.

To evaluate the feature augmentation, we sample 1,000 training samples from the Adult dataset, extract the sensitive features with the trained model, and generate the augmented features. The trained gender classifier, whose prediction accuracy is 86.8% when using the original sensitive features, predicts opposite labels for all the augmented features. This means the augmentation policy successfully generates the features corresponding to the opposite sensitive attribute.

To evaluate the prediction accuracy of the imputation model on the augmented samples, we select all the pairs of training samples (208 pairs in total) with the same other attributes but different sensitive attributes and different labels. Intuitively, an accurate imputation model should predict the opposite label for the augmented samples since every augmented sample has a real sample with an opposite label corresponding to it. The imputation model predicts the opposite label on 258 augmented samples (out of 416 samples), which means the prediction accuracy of the task model can be improved with label calibration.

To better understand the training process of Fair-CDA, we

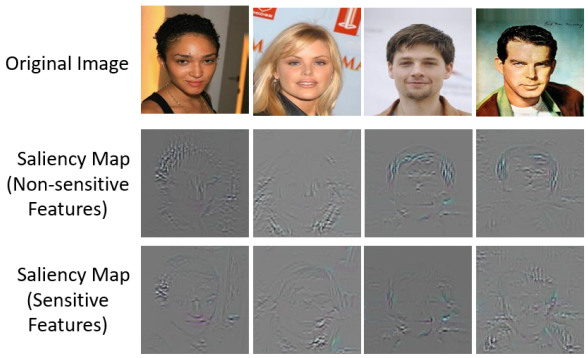


Figure 4: Saliency map on the wavy hair recognition task. The saliency maps of sensitive features focus more on the whole face, while those of non-sensitive features focus more on the hair of a man/woman.

plot the trend of loss and performance during two stages on the Adult dataset, as shown in Figure 3. Stage 1 stands for the process of the model trained with original data, while Stage 2 stands for that of the model trained with augmented data. At the beginning of Stage 1, the model is initialized with random parameters (without pre-training). The mean value of the output of both groups tends to be very close, resulting in low fairness disparity. As the training process goes on, the loss of Fair-CDA converges gradually with the AP rising to a high point. Both DP and EO reach a high value (which means poor fairness). In Stage 2, the AP remains stable, the loss fluctuates at the beginning, and finally drops to a low point. Both DP and EO reach a low value.

Results on CelebA. To illustrate each model's ability for vision tasks, we choose smiling, wavy hair, and attractive to form three binary classification tasks. As shown in Figure 5, Fair-CDA achieves SOTA performance followed by two mixup methods. It is worth mentioning that the DP and EO gap of these methods on the smiling recognition task is smaller compared with other tasks, which is a relatively fair scenario, but Fair-CDA can still improve the fairness. Also, Fair-CDA is the only method that achieves considerable accuracy given high fairness requirements on both tasks.

To visualize the effect of feature decomposition, we adopt deep neural network interpretability methods in (Adebayo et al. 2018). We draw the saliency map on the wavy hair recognition task, as shown in Figure 4. Sensitive features are those strongly related to gender, while non-sensitive features are those strongly related to wavy hair. It can be seen that the saliency maps of sensitive features focus more on the whole face, while those of non-sensitive features focus more on the hair of a man/woman.

Additionally, we evaluate Fair-CDA on more sensitive features on CelebA dataset. We implement Fair-CDA on the same task as that in (Creager et al. 2019) (CelebA Heavy-Makeup recognition task) to compare our method with two VAE methods. Noted that the fairness metric is demographic parity and the performance metric is accuracy in this setting. As shown in Table 2, Fair-CDA outperforms FFVAE and β-VAE on CelebA Heavy-Makeup recognition task considering three different sensitive attributes.

	Male Δ_{DP}/Acc	Chubby Δ_{DP}/Acc	Eyeglasses Δ_{DP}/Acc
β-VAE	0.330/0.712	0.202/0.732	0.250/0.715
	0.400/0.725	0.220/0.740	0.280/0.735
FFVAE	0.330/0.730	0.202/0.748	0.250/0.725
	0.400/0.752	0.400/0.825	0.400/0.824
Fair-CDA	0.234/0.733	0.184/0.816	0.217/0.814
	0.369/0.836	0.197/0.825	0.245/0.824

Table 2: Results on CelebA dataset. Compared with two VAE methods, Fair-CDA improves the fairness measurement Δ_{DP} and accuracy significantly.

Results on MovieLens dataset. Recommendation, a common scenario of machine learning, poses unique challenges for applying fairness and non-discrimination concepts. We choose the rating recognition task on MovieLens to evaluate different fair methods. Similar to the trends on the Adult dataset, Fair-CDA achieves SOTA performance followed by Fair Mixup and GapReg, as shown in Figure 6. AdvDebias achieves better fairness than ERM accompanied by severe accuracy degradation. In addition, Fair-CDA can reach the smallest Δ_{DP} and Δ_{EO} among all the methods, which shows its superior ability to obtain the group fairness.

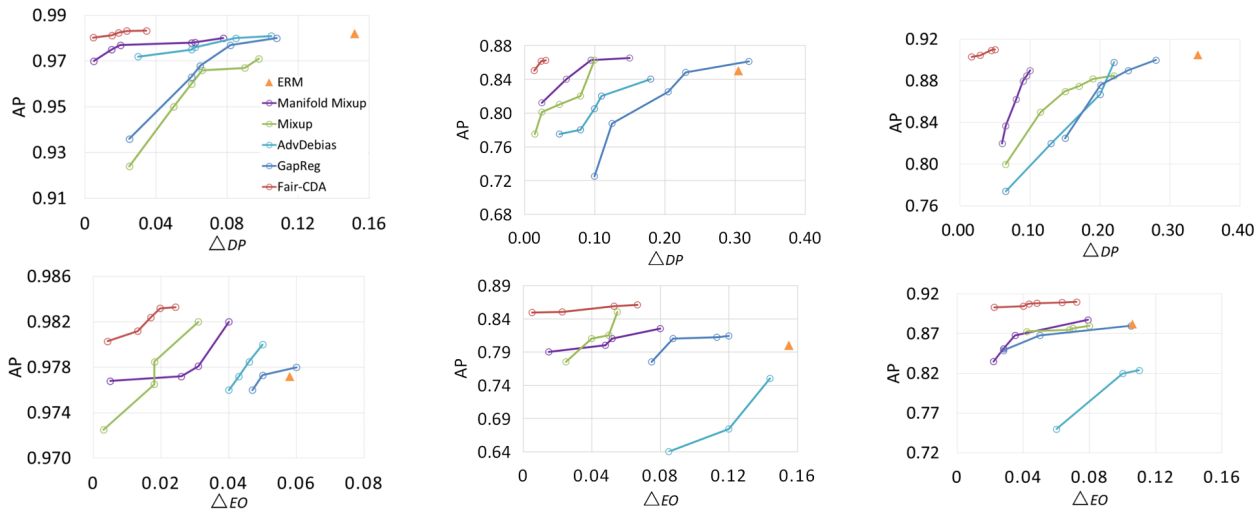
Ablation Studies

Without imputation model. To examine whether the imputation model contributes to performance, we train Fair-CDA without imputation model (Fair-CDA (no IM)) on MovieLens dataset and plot the Pareto Front in Figure 7. We can see the Pareto Front of Fair-CDA dominates that of Fair-CDA (no IM) for both DP and EO. Without an imputation model, Fair-CDA can still achieve good fairness but suffer a little accuracy loss.

Sample generating at the attribute level. To illustrate the effectiveness of feature decomposition, we use a naive way to generate the flip sample. Simply flipping the value of the sensitive attribute with a specific probability during the training phase, we can get the results of model training with different data distributions. We name the method as Attribute-Level. By setting different probabilities, we can get the Pareto Front, as shown in Figure 7. Compared with ERM, Attribute-Level can mitigate the unfairness to some extent, while it can not solve the problem mentioned earlier: other variables correlated with sensitive attributes can serve as a source for unfairness.

Experiments on Product.

To further validate our algorithm in realistic scenarios, we deploy Fair-CDA in an online course recommender system. There are nearly 100,000 users and more than 12,000 courses developed by more than 100 different suppliers, nearly 50% of the courses coming from top 5% suppliers. Thus we consider supplier as the sensitive attribute to evaluate fairness. Similar to the previous setting on MovieLens, we divide the suppliers into the majority and minor-



((a)) Smiling ((b)) Wavy Hair ((c)) Attractive

Figure 5: CelebA. The trade-off between AP and Δ_{DP} / Δ_{EO}. Fair-CDA outperforms other methods across tasks.

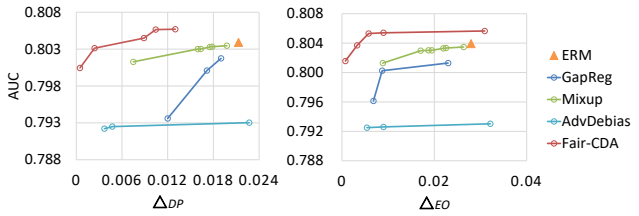


Figure 6: MovieLens. The trade-off between AP and Δ_{DP} / Δ_{EO}. Fair-CDA can reach the smallest Δ_{DP} / Δ_{EO} among all the methods without obvious accuracy degradation.

Method	AUC	Δ_{EO}	Top-10 Recall
LightGCN (Baseline)	0.9503	0.0448	0.1116
Fair-CDA	0.9679	0.0227	0.1328

Table 3: Offline results on a product dataset from an online course recommender system.

ity groups according to the number of courses developed by the suppliers. The top 5% suppliers who provide nearly 50% of the courses are regarded as the majority supplier and the remaining suppliers are regarded as the minority supplier. In this scenario, we choose Equalized Odds as the fairness measurement since it has been shown that demographic parity causes a loss in the utility and infringes individual fairness (Singh and Joachims 2018), and we adopt AUC and Top-10 Recall as the offline accuracy evaluation. We use LightGCN (He et al. 2020) as the backbone network and compare Fair-CDA with the original LightGCN method. The results are shown in Table 3. Fair-CDA achieves better performance on both accuracy and fairness measurements than the baseline method.

Inspired by the performance of offline evaluation, we implement and deploy Fair-CDA in the production environment and verify its effectiveness through a consecutive online A/B test. We split the users into two groups uniformly, each of which has an average of 3000 users every week. The first group gets courses recommended by the baseline model, and the Fair-CDA generates recommendations for the other group. The two models are updated daily. After a 5-week online A/B test, the Fair-CDA is consistently superior than the baseline model, with an average Click Through Rate (CTR) improvement of 6.5%. During the online A/B test, our method increases the diversity of recommended courses and enhances group fairness, resulting in a higher CTR.

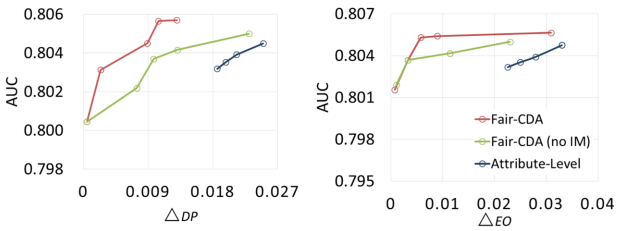


Figure 7: Ablation studies on MovieLens dataset. Fair-CDA without imputation model (Fair-CDA (no IM)) can still satisfy the fairness requirement but suffer an accuracy loss. By generating the samples with opposite sensitive attributes (Attribute-Level), the unfairness can hardly be decreased.

Conclusions

We propose *Fair-CDA* to counter the unfairness problem via feature decomposition and data augmentation. Fair-CDA improves fairness and minimizes the impact on accuracy. We experimentally compare our method with other state-of-the-art fairness methods on various benchmarks and show that Fair-CDA significantly outperforms the other methods in all the experimental settings.

Acknowledgements

This work was partially supported by JCYJ20220530143600001, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen HK S&T Cooperation Zone, by the National Key R&D Program of China with grant No.2018YFB1800800, by SGDX20211123112401002, by Shenzhen Outstanding Talents Training Fund, by Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), by the NSFC 61931024&8192 2046, by NSFC-Youth 62106154, by zelixir biotechnology company Fund, by Tencent Open Fund, and by ITSO at CUHKSZ. Chuanlong Xie was partially supported by NSFC No.12201048 and the Interdisciplinary Intelligence SuperComputer Center of Beijing Normal University at Zhuhai.

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*.

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, 60–69. PMLR.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization. *arXiv:1907.02893*.

Bai, H.; Sun, R.; Hong, L.; Zhou, F.; Ye, N.; Ye, H.-J.; Chan, S.-H. G.; and Li, Z. 2020. DecAug: Out-of-Distribution Generalization via Decomposed Feature Representation and Semantic Augmentation. *arXiv preprint arXiv:2012.09382*.

Barocas, S.; Hardt, M.; and Narayanan, A. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

Beutel, A.; Chen, J.; Doshi, T.; Qian, H.; Wei, L.; Wu, Y.; Heldt, L.; Zhao, Z.; Hong, L.; Chi, E. H.; et al. 2019. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2212–2220.

Black, E.; Yeom, S.; and Fredrikson, M. 2020. Fliptest: fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 111–121.

Bose, A.; and Hamilton, W. 2019. Compositional fairness constraints for graph embeddings. In *International Conference on Machine Learning*, 715–724. PMLR.

Calmon, F. P.; Wei, D.; Vinzamuri, B.; Ramamurthy, K. N.; and Varshney, K. R. 2017. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 3995–4004.

Chuang, C.-Y.; and Mroueh, Y. 2021. Fair Mixup: Fairness via Interpolation. In *International Conference on Learning Representations*.

Cotter, A.; Gupta, M.; Jiang, H.; Srebro, N.; Sridharan, K.; Wang, S.; Woodworth, B.; and You, S. 2019. Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In *International Conference on Machine Learning*, 1397–1405. PMLR.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly Fair Representation Learning by Disentanglement. In *International conference on machine learning*, 1436–1445. PMLR.

Dou, Q.; Castro, D. C.; Kamnitsas, K.; and Glocker, B. 2019. Domain Generalization via Model-Agnostic Learning of Semantic Features. In *Advances in Neural Information Processing Systems*.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml. Accessed: 2021-03-01.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Edwards, H.; and Storkey, A. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*.

Ekstrand, M. D.; Das, A.; Burke, R.; and Diaz, F. 2021. Fairness and Discrimination in Information Access Systems. *arXiv preprint arXiv:2105.05779*.

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

Feng, R.; Yang, Y.; Lyu, Y.; Tan, C.; Sun, Y.; and Wang, C. 2019. Learning fair representations via an adversarial framework. *arXiv preprint arXiv:1904.13341*.

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

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29: 3315–3323.

Harper, F. M.; and Konstan, J. A. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4): 1–19.

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

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, 639–648.

Jha, A.; Vinzamuri, B.; and Reddy, C. K. 2021. Fair Representation Learning using Interpolation Enabled Disentanglement. *arXiv preprint arXiv:2108.00295*.

Joachims, T.; and Swaminathan, A. 2016. Counterfactual Evaluation and Learning for Search, Recommendation and

Ad Placement. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, 1199–1201.

Kamiran, F.; and Calders, T. 2009. Classifying without discriminating. In *International Conference on Computer*.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1): 1–33.

Khademi, A.; Lee, S.; Foley, D.; and Honavar, V. 2019. Fairness in algorithmic decision making: An excursion through the lens of causality. In *The World Wide Web Conference*, 2907–2914.

Kuang, K.; Cui, P.; Athey, S.; Xiong, R.; and Li, B. 2018. Stable Prediction across Unknown Environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1617–1626.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *Advances in neural information processing systems*, 30.

Lahoti, P.; Gummadi, K. P.; and Weikum, G. 2019. ifair: Learning individu ally fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 1334–1345. IEEE.

Li, Y.; Xu, K.; Lai, R.; and Gu, L. 2022. Towards an Effective Orthogonal Dictionary Convolution Strategy. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, volume 36, 1473–1481.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2018. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15(2018): 11.

Locatello, F.; Abbati, G.; Rainforth, T.; Bauer, S.; Schölkopf, B.; and Bachem, O. 2019. On the Fairness of Disentangled Representations. *Advances in Neural Information Processing Systems*, 32: 14611–14624.

Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. 2015. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, 3384–3393. PMLR.

McNamara, D.; Ong, C. S.; and Williamson, R. C. 2017. Provably fair representations. *arXiv preprint arXiv:1710.04394*.

McNamara, D.; Ong, C. S.; and Williamson, R. C. 2019. Costs and benefits of fair representation learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 263–270.

Morik, M.; Singh, A.; Hong, J.; and Joachims, T. 2020. Controlling Fairness and Bias in Dynamic Learning-to-Rank. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Park, S.; Hwang, S.; Kim, D.; and Byun, H. 2021. Learning Disentangled Representation for Fair Facial Attribute Classification via Fairness-aware Information Alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2403–2411.

Quadrianto, N.; Sharmanska, V.; and Thomas, O. 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8227–8236.

Sarhan, M. H.; Navab, N.; Eslami, A.; and Albarqouni, S. 2020. Fairness by learning orthogonal disentangled representations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, 746–761. Springer.

Singh, A.; and Joachims, T. 2018. Fairness of exposure in rankings. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2219–2228.

Song, J.; Kalluri, P.; Grover, A.; Zhao, S.; and Ermon, S. 2019. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 2164–2173. PMLR.

Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; and Bengio, Y. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, 6438–6447. PMLR.

Wang, R.; Fu, B.; Fu, G.; and Wang, M. 2017. Deep & cross network for ad click predictions. In *Proceedings of the ADKDD'17*, 1–7.

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

Zafar, M. B.; Valera, I.; Rogriguez, M. G.; and Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, 962–970. PMLR.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *International Conference on Machine Learning*, 325–333.

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

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *International Conference on Learning Representations*.

Zhang, L.; Deng, Z.; Kawaguchi, K.; Ghorbani, A.; and Zou, J. 2021. How Does Mixup Help With Robustness and Generalization? In *International Conference on Learning Representations*.

Zhou, N.; Zhang, Z.; Nair, V. N.; Singhal, H.; Chen, J.; and Sudjianto, A. 2021. Bias, Fairness, and Accountability with AI and ML Algorithms. *arXiv preprint arXiv:2105.06558*.