DualFair: Fair Representation Learning at Both Group and Individual Levels via Contrastive Self-supervision

Sungwon Han
KAIST
Daejeon, South Korea
lion4151@kaist.ac.kr

Seungeon Lee
KAIST
Daejeon, South Korea
archon159@kaist.ac.kr

Fangzhao Wu
Microsoft Research Asia
Beijing, China
wufangzhao@gmail.com

Sundong Kim
GIST
Gwangju, South Korea
sundong@gist.ac.kr

Chuhan Wu
Tsinghua University
Beijing, China
wuchuhan15@gmail.com

Xiting Wang
Microsoft Research Asia
Beijing, China
xitwan@microsoft.com

Xing Xie
Microsoft Research Asia
Beijing, China
xingx@microsoft.com

Meeyoung Cha
IBS & KAIST
Daejeon, South Korea
mcha@ibs.re.kr

ABSTRACT

Algorithmic fairness has become an important machine learning problem, especially for mission-critical Web applications. This work presents a self-supervised model, called DualFair, that can debias sensitive attributes like gender and race from learned representations. Unlike existing models that target a single type of fairness, our model jointly optimizes for two fairness criteria—*group fairness* and *counterfactual fairness*—and hence makes fairer predictions at both the group and individual levels. Our model uses contrastive loss to generate embeddings that are indistinguishable for each protected group, while forcing the embeddings of counterfactual pairs to be similar. It then uses a self-knowledge distillation method to maintain the quality of representation for the downstream tasks. Extensive analysis over multiple datasets confirms the model's validity and further shows the synergy of jointly addressing two fairness criteria, suggesting the model's potential value in fair intelligent Web applications.

CCS CONCEPTS

• **Computing methodologies → Philosophical/theoretical foundations of artificial intelligence**; **Machine learning**.

KEYWORDS

Fair representation learning, Unsupervised learning, Contrastive learning, Group fairness, Counterfactual fairness.

1 INTRODUCTION

Machine learning techniques are being used in many mission-critical real-world Web applications, such as recommendation [1], hiring [25], and advertisement of profiles [41]. Many tasks are subject to stereotypes or societal biases embedded in data, leading models to treat individuals of particular attributes unfairly. Notable examples include facial recognition models that fail to recognize people with darker skin [9] or recruiting models that favor men over women candidates with comparable work experience [29]. Lacking fairness in these tasks aggravates social inequity and even harms individuals and society. As a response, a growing number of studies are dedicated to algorithmic fairness [17, 48].

Fairness is also important in representation learning. Fair representation learning is useful when data needs to be shared without any prior knowledge of the downstream task. Previous research has focused on finding data representations that apply to diverse tasks while considering fairness. For example, LAFTR [35] and VFAE [34] introduce an adversarial concept, and studies like [40] and [43] maximize the conditional mutual information with fairness constraints to reduce the potential bias. These approaches generate fairer embeddings, albeit at the cost of performance degradation in downstream tasks. They also consider only group-level fairness, missing out on individual-level fairness [6].

We propose DualFair, a self-supervised learning method that debiases sensitive information through fairness-aware contrastive learning while preserving rich expressiveness (i.e., representation quality) through self-knowledge distillation. DualFair can learn data representations satisfying two types of fairness, i.e., *group fairness* and *counterfactual fairness*. The former (a.k.a. demographic parity) requires every protected group be treated in the same way as any advantaged group, and the latter requires the model to treat individuals in a counterfactual relationship (i.e., those who share similar traits except for the sensitive attribute) alike [32]. Counterfactual fairness is a type of fairness defined at the individual level. It removes bias from sensitive information by comparing against synthetic individuals from the counterfactual world.

Because the model is unaware of downstream tasks during training, adding fairness-related regularization terms (e.g., minimizing demographic parity from model predictions [26]) or fairness constraints (e.g., limiting the prediction difference among groups below the threshold [51]) as in other research is not feasible. Instead, we

propose the following alternative goals for our loss design, which are illustrated in Figure 1:

(1) **Group fairness**: Data points from every sensitive group have the same distribution across the embedding space; thus, their group membership cannot be identified.

(2) **Counterfactual fairness**: Data points from those in a counterfactual relationship are located close in the embedding space. This means the embedding distance between an individual and its counterfactual version should be minimized.

These goals can apply universally to any downstream task. By making sensitive attributes indistinguishable in the embedding space, a downstream classifier cannot determine which data point belongs to the protected group and consequently produce unbiased predictions. The same applies to embeddings of counterfactual pairs.

To implement these fairness objectives, we first propose a cyclic variational autoencoder (C-VAE) model to generate counterfactual samples (Sec. 3.2). This sample generation task is non-trivial due to the high correlation between data features and sensitive attributes. Next is to create fairness-aware embeddings from the input data. We employ contrastive learning, which learns representations based on the similarity between instances by modifying the selection of positive and negative samples. We select negative samples from the same protected group and select positive samples from counterfactual versions (Sec. 3.3). In addition, we use self-knowledge distillation to extract semantic features and enforce consistent embeddings between the original and its augmentation to maintain a high representation quality (Sec. 3.4). Experiments demonstrate that the proposed framework can generate data embeddings that satisfy both fairness criteria while preserving prediction accuracy for a variety of downstream tasks. The main contributions of this paper are summarized below.

- We propose a **self-supervised representation learning framework (DualFair)** that simultaneously debiases sensitive attributes at both group and individual levels.
- We introduce the C-VAE model to generate counterfactual samples and propose **fairness-aware contrastive loss** to meet the two fairness criteria jointly.
- We design the **self-knowledge distillation loss** to maintain representation quality by minimizing the embedding discrepancy between original and perturbed instances.
- Experimental results on six real-world datasets confirm that DualFair generates a fair embedding for sensitive attributes while maintaining high representation quality. The ablation study further shows a synergistic effect of the two fairness criteria.

Codes are released at a GitHub repository.[1]

2 RELATED WORKS

2.1 Fairness in Machine Learning

Fairness is a conceptual term, and cannot be measured in a straightforward manner. Instead, there are several criteria to observe it from different perspectives: unawareness, group fairness, and individual fairness [20, 44]. Removing sensitive attributes from data is a simple way to achieve fairness through unawareness [12]. However, it can
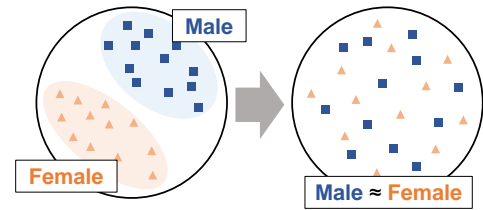
(a) **Group fairness: Embeddings of different member groups (e.g., by gender) should not be distinguished. This makes unbiased predictions possible at the group level.**
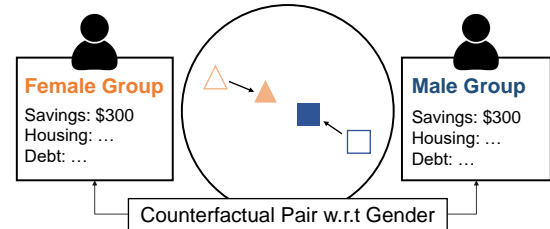
(b) **Counterfactual fairness (i.e., individual-level fairness): Individuals of similar traits should have similar embeddings irrespective of their sensitive attributes. This makes unbiased predictions possible at the individual level.**

Figure 1: Illustration of two fairness criteria.

be brittle if some hidden features are highly correlated with sensitive attributes. Group fairness states that subjects in different groups (e.g., gender) should have an equal probability of being assigned to the predicted class [15, 20]. Demographic parity and equalized odds are two measures of group fairness [23, 50]. Individual fairness is a fine-grained criterion that treats similar individuals as similarly as possible [18]. Counterfactual fairness is one alternative to individual fairness that assumes a counterfactual sample by flipping the sensitive attributes and treats it similarly to the original one [39].

Researchers also focus on diverse learning steps to ensure fairness. Elazar et al. [19] focus on fairness of the input dataset by noting that a fair decision is made with the model trained with an unbiased dataset. However, Wang et al. [46] demonstrate that input-wise fairness does not entirely support fair decisions in large-scale datasets. Therefore, post-processing techniques are proposed to secure fairness, such as hiding the sensitive attribute information from trained representations by null space projection [37] or identifying the subspace of sensitive attributes [7].

With the advent of representation learning, recent self-supervised learning approaches aim to produce a fair representation of individual instances without knowing downstream tasks (i.e., treatment-level fairness) [31]. These fair representation learning approaches add fairness-related objectives in training steps, or apply adversarial learning objectives to obtain fair representations [5, 22, 33, 52]. For example, LAFTR [35] employs a discriminator that detects sensitive attribute information, while generators make indistinguishable representations against discriminators. VFAE [34] proposes a variational autoencoder with regularization using Maximum Mean Discrepancy (MMD) to learn fair representations.
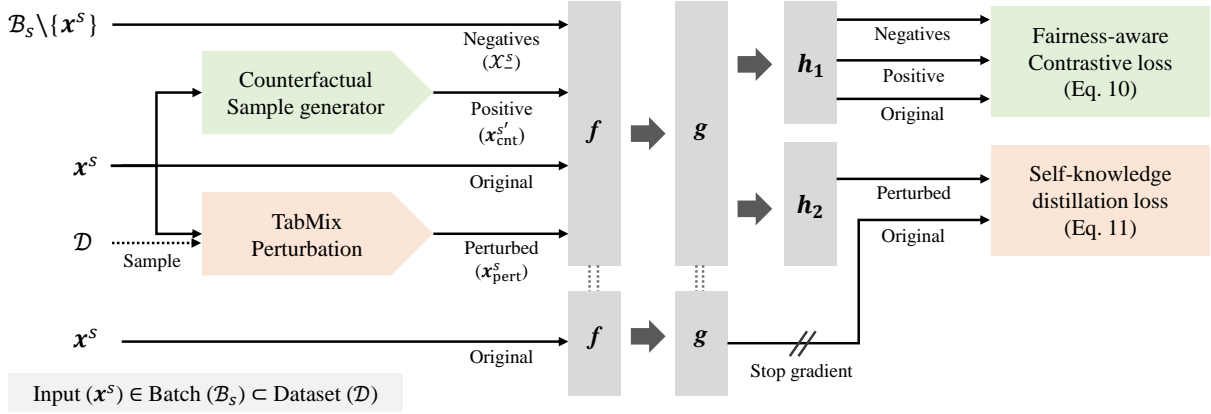
---

[1] https://github.com/Sungwon-Han/DualFair

Figure 2: The overall architecture of DualFair, where f, g, (h_1, h_2) denote the backbone network, the projection head, and two prediction heads, respectively. DualFair aims to achieve both fairness and representation quality by jointly optimizing the fairness-aware contrastive loss and the self-knowledge distillation loss.

Song et al. [40] proposes a user-centric approach that allows users to control the level of fairness and maximize the expressiveness of representations in the form of conditional mutual information with the preset fairness threshold. Tsai et al. [43] aim to optimize the same objective but maximize the lower bound of conditional mutual information via InfoNCE objective instead. However, these methods only consider the single group fairness objective and often lose the expressiveness during the training [10, 11]. Specifically, group fairness helps achieve anti-discrimination for protected groups, but individual justice is not guaranteed [6]. Our research objective is to implement multiple algorithmic fairness concepts on a single deep model, a topic that is explored less in the literature. We seek to achieve both group fairness at a coarse-grained level and counterfactual fairness at an individual level while preserving a high level of representational quality.

2.2 Contrastive Self-Supervised Learning

The fundamental idea of contrastive learning is to minimize the distance between similar (i.e., positive) instances while maximizing the distance among dissimilar (i.e., negative) instances [13, 24]. Sim-CLR [13] utilizes augmented images as positives while the other images in the same batch as negatives. Maintaining a similar contrastive concept, MoCo [24] exploits a momentum encoder and proposes a dynamic dictionary with a queue to handle negative samples efficiently in both performance and memory perspectives. InfoNCE loss [36] is often used in contrastive learning. Minimizing this loss increases mutual information between positive pairs so that the model can extract the consistent features between the original and augmented samples.

3 METHODOLOGY

3.1 Overview

We present DualFair, a self-supervised learning method that ensures both group and counterfactual fairness criteria while maintaining high representation quality with the design of two special losses. We use a pictorial sketch in Figure 2 to describe them.

First is the **fairness-aware contrastive loss**, which treats individuals in *counterfactual relationships alike* (i.e., counterfactual fairness) and ensures *non-distinguishable embeddings* over sensitive attributes (i.e., group fairness). A key to this loss is the design of a sample generator that produces a counterfactual version $\mathbf{x}_{\text{cnt}}^{s'}$ of an input item $\mathbf{x}^s$ by maintaining its latent characteristics but flipping the sensitive attribute $s \rightarrow s'$ (Sec. 3.2). In a job interview, for instance, this corresponds to a hypothetical decision if the applicant's gender were to change, with all other latent traits, such as education and work experience, remaining unchanged [32]. Given a dataset $\mathcal{D}$ with sensitive attributes S like gender, the fairness-aware contrastive loss is defined on the batch $\mathcal{B}_s$, whose data instances share the same sensitive attribute $s \in S$ (e.g., gender is female). The loss maximizes agreements between the original item and its counterfactual version generated by our generator (i.e., ensuring counterfactual fairness) and minimizes agreements between items with the same sensitive attribute (i.e., ensuring group fairness) (Sec. 3.3).

Second is the **self-knowledge distillation loss** for extracting semantic information from the instance and maintaining the performance of downstream tasks (Sec. 3.4). Consider a Siamese network that has two heads: one for the student branch and the other for the teacher branch. We impose the perturbed instance's embedding from the student branch to be similar to that of the original instance from the teacher branch. The rationale behind this treatment is that a perturbation that does not largely deform the original content should not change the learned representation. We introduce our unique perturbation module TabMix for an efficient self-knowledge distillation process.

3.2 Counterfactual Sample Generation

Our sample generator creates data following the definition of a counterfactual relationship in [32]. Let us denote three variables V, U, R, where V is a set of observable variables, including the sensitive attribute S and other attributes X. U is a set of latent variables
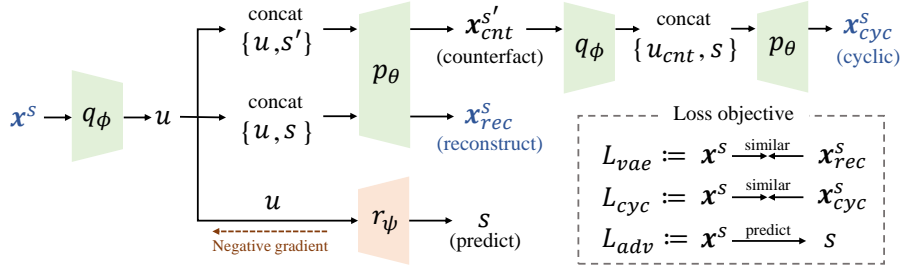
Figure 3: Design of our counterfactual sample generator. Adversarial training over VAE eliminates sensitive information s from u. Cyclic consistency loss makes training more stable by ensuring the double-flipped sample (x^s_{cyc}) to be same as the original (x^s).
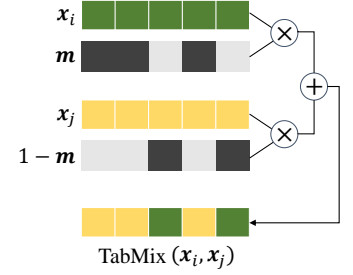

Figure 4: Illustration of TabMix operation for self-knowledge distillation (m denotes a mask).

independent of S or V. R is a set of functions representing causal relationships from U to V, or between V. Considering a causal model with (V, U, R), counterfactual inference works in three steps:

(1) Calculate the posterior distribution of latent variables U from the input instance.
(2) Choose a target sensitive attribute $s \in S$ (e.g., gender) and reformulate a set of functions R assuming the sensitive attribute $s \in S$ is fixed in the causal graph.
(3) Infer V from U and $s \in S$ using re-formulated R.

Extracting the latent variable U is non-trivial because it should explain observable variables sufficiently without revealing information about sensitive attributes.

For this task, we introduce the cyclic variational autoencoder (C-VAE) depicted in Figure 3. Our counterfactual sample generator is trained in the form of $p(x^s, \mathbf{u}, s) = p(\mathbf{u})p(s)p(x^s|\mathbf{u}, s)$, where $\mathbf{u}$ and s represent the latent variable and the sensitive attribute in the input x^s. $p(\mathbf{u})$ and $p(s)$ represent the prior distribution of the latent variable $\mathbf{u}$ from the encoder and the sensitive attribute s. $p(\mathbf{u})$ is assumed to follow a standard normal distribution [28]. Then $p(x^s|\mathbf{u}, s)$ is a likelihood function from the decoder that reconstructs the input sample x^s from $\mathbf{u}$ and s. The encoder and decoder networks are denoted as q_ϕ and p_θ. We formulate $L_{\text{vae}}(x^s)$ from the variational lower bound of $\log p(x)$ as follows:

$$L_{\text{vae}}(x^s) = \text{KL}(q_\phi(\mathbf{u}|x^s)||p(\mathbf{u})) - \mathbb{E}_{q_\phi(\mathbf{u}|x^s)}[\log p_\theta(x^s|\mathbf{u}, s)]. \quad (1)$$

Latent variable $\mathbf{u}$ from the encoder q_ϕ and sensitive attribute s are not guaranteed to be independent even if they correspond to U and S in the causal model that are independent. To ensure independence, we introduce an additional adversarial objective and lead the model to eliminate unnecessary sensitive information from $\mathbf{u}$. A discriminator r_ψ is trained with a cross-entropy loss to predict the sensitive attribute s from the latent representation $\mathbf{u}$:

$$L_{\text{adv}}(x^s) = -\mathbb{E}_{q_\phi(\mathbf{u}|x^s)}[\log r_\psi(s|\mathbf{u})]. \quad (2)$$

We negate the cross-entropy loss to train the sample generator to prevent the discriminator r_ψ from predicting sensitive attributes.

We further introduce the property of **cyclic consistency** to enhance training stability and improve generation quality. Our loss ensures that the counterfactual-counterfactual sample (i.e., double-flipped sample created by our generator) is similar to the original

item. We produce a counterfactual sample $x^{s'}_{cnt}$ of the input instance x^s, repeating the three steps below:

(1) Compute a latent representation $\mathbf{u}$ from the encoder q_ϕ.
(2) Choose the target sensitive attribute $s' (\neq s)$.
(3) Reconstruct the counterfactual sample from the decoder p_θ given the latent variable $\mathbf{u}$ and the target sensitive attribute s'.

The reconstructed sample $x^{s'}_{cnt}$ goes back to the generator to produce a counterfactual-counterfactual (or double-flipped) sample x^s_{cyc} that shares the identical sensitive attribute with the original input. Our cyclic consistency loss in C-VAE maximizes the likelihood between two instances; x^s, x^s_{cyc} (Eq. 3).

$$convert(x^{s'}_{cnt}|x^s, s') = p_\theta(x^{s'}_{cnt}|q_\phi(\mathbf{u}|x^s), s'),$$

$$L_{\text{cyc}}(x^s) = -\mathbb{E}_{convert(x^{s'}_{cnt}|x^s, s')}[\log convert(x^s|x^{s'}_{cnt}, s)] \quad (3)$$

The total loss to train the sample generator is as follows:

$$L_{\text{c-vae}}(x^s) = L_{\text{vae}}(x^s) - L_{\text{adv}}(x^s) + L_{\text{cyc}}(x^s). \quad (4)$$

3.3 Fairness-aware Contrastive Learning

The next objective, given the learned sample generator, is to train an encoder that satisfies both group and counterfactual fairness. We use contrastive learning for this task. Let us denote an input query as $\mathbf{x} \in \mathbb{R}^d$, and a positive sample and a set of negative samples for contrastive learning as $\mathbf{x}_+$ and $\mathcal{X}_-$, respectively. Typically, positive samples and negative samples are determined by predefined rules. For example, a rule can set a single data instance and its augmented versions as positive, whereas set all others as negative (see description in §2.2). In this work, we define InfoNCE loss in contrastive learning to train the model F (i.e., $F := h_1 \circ g \circ f$ in Figure 2) as follows:

$$L_c(\mathbf{x}, \mathbf{x}_+, \mathcal{X}_-) = -\log \frac{\exp(\text{sim}(F(\mathbf{x}), F(\mathbf{x}_+))/\tau)}{\sum_{\mathbf{x}' \in \{\mathbf{x}_+\} \cup \mathcal{X}_-} \exp(\text{sim}(F(\mathbf{x}), F(\mathbf{x}'))/\tau)} \quad (5)$$

$$= -\text{sim}(F(\mathbf{x}), F(\mathbf{x}_+))/\tau$$

$$+ \log \sum_{\mathbf{x}' \in \{\mathbf{x}_+\} \cup \mathcal{X}_-} \exp(\text{sim}(F(\mathbf{x}), F(\mathbf{x}'))/\tau) \quad (6)$$

$$= -L_{\text{align}}(\mathbf{x}, \mathbf{x}_+) + L_{\text{distribution}}(\mathbf{x}, \mathbf{x}_+, \mathcal{X}_-), \quad (7)$$

where $\text{sim}(\cdot)$ is the function to measure two embeddings' similarity, and τ is the temperature parameter.

This contrastive loss in Eq. 7 comprises two terms. First, the alignment loss (L_{align}) encourages the embedding positions of positive

pairs to be placed closer. Second, in contrast, the distribution loss ($L_{\text{distribution}}$) matches all instances' embeddings into the prior distribution with a high entropy value. Our model uses the generalized contrastive objective proposed in [45], which supports the diverse choice of prior distributions by introducing the optimal transport theory [8] and by changing the distribution loss with Sliced Wasserstein Distance (SWD) [30]. Given the prior distribution over the embedding space $\mathcal{Z}_{prior}$ and the set of embeddings $\tilde{\mathcal{Z}} = \{F(\mathbf{x})|\mathbf{x} \in \mathcal{X}_-\}$, the loss is formulated as the following equation:

$$L_{\text{gen-}c}(\mathbf{x}, \mathbf{x}_+, \mathcal{X}_-) = -L_{\text{align}}(\mathbf{x}, \mathbf{x}_+) + \text{SWD}(\tilde{\mathcal{Z}}, \mathcal{Z}_{prior}). \quad (8)$$

We modify the contrastive loss to jointly meet two fairness criteria. Assume that all instances in $\mathcal{X}_-$ are sampled to have the same sensitive attribute s (e.g., gender = female). Then minimizing the distribution loss on $\mathcal{X}_-$ (i.e., $\text{SWD}(\tilde{\mathcal{Z}}, \mathcal{Z}_{prior})$ in Eq. 8) will cause the sensitive group s in embeddings to match the predefined prior distribution $\mathcal{Z}_{prior}$. By iterating this process for every sensitive group (e.g., female and male), our model can produce embeddings of groups that follow the same distribution, $\mathcal{Z}_{prior}$. As a result, embeddings become no longer distinguishable by sensitive attributes. Given a batched set of instances $\mathcal{B}_s$ and an input $\mathbf{x}^s$, negative samples are defined as follows:

$$\mathcal{X}_-^s = \{\mathbf{x}|\mathbf{x} \in \mathcal{B}_s \setminus \{\mathbf{x}^s\}\}. \quad (9)$$

To ensure counterfactual fairness, our model also considers a counterfactual version of the sample as positive in contrastive learning. Given an input sample $\mathbf{x}^s$, we flip the sensitive attribute $s \to s'$ (e.g., female to male) and generate the counterfactual sample $\mathbf{x}_{\text{cnt}}^{s'}$ from our C-VAE based sample generator. The alignment loss in Eq. 8 then minimizes the embedding discrepancy between the original and counterfactual data instances. From the positive sample and the set of negative samples ($\mathbf{x}_{\text{cnt}}^{s'}, \mathcal{X}_-^s$), the fairness-aware contrastive loss for the given input $\mathbf{x}^s$ is defined as follows:

$$L_{\text{fair-cl}}(\mathbf{x}^s) = L_{\text{gen-}c}(\mathbf{x}^s, \mathbf{x}_{\text{cnt}}^{s'}, \mathcal{X}_-^s). \quad (10)$$

We train the embedding on top of the Euclidean space with Gaussian prior $\mathcal{Z}_{prior}$. This is different from other contrastive approaches where embeddings are learned over the L2-normalized space [13, 24]. L2-normalized space used in those approaches does not account for the norm of embeddings during training, which can also be an important clue for sensitive attributes in downstream tasks. However, our choice of the Euclidean space regularizes the norm distribution and removes its dependency on sensitive attributes. The alignment loss L_{align} is then defined with a negative Euclidean distance as a similarity measure between the original and positive instance.

3.4 Self-knowledge Distillation

The final objective is maintaining the representation quality. We design the self-knowledge distillation loss to reduce the embedding discrepancy between the original and perturbed instances. Inspired by the original literature [21], we present a Siamese network with two different heads, where each head becomes the student and teacher branches. Then, the model is trained through a prediction task so that the original instance from the teacher branch is highly predictive of the perturbed instances from the student branch. This process is called *self-knowledge distillation*, since the knowledge extracted from the teacher branch is progressively transferred back

to the student branch [27, 42]. To prevent the model from collapsing into a naive solution (e.g., representation becomes constant for every instance), we let the architecture of student and teacher be asymmetric, restricting the gradient flows of the teacher branch [14, 21] (see the bottom part of Figure 2). Let us denote f, g, and h_2 as the backbone network, the projection head, and the prediction head, respectively. The self-knowledge distillation loss is defined as follows:

$$\mathbf{p}_{\text{student}} = h_2 \circ g \circ f(\mathbf{x}_{\text{pert}}^s), \quad \mathbf{z}_{\text{teacher}} = sg \circ g \circ f(\mathbf{x}^s)$$

$$L_{\text{self-kd}}(\mathbf{x}^s) = -\frac{\mathbf{p}_{\text{student}}}{||\mathbf{p}_{\text{student}}||_2} \cdot \frac{\mathbf{z}_{\text{teacher}}}{||\mathbf{z}_{\text{teacher}}||_2}, \quad (11)$$

where $\mathbf{x}_{\text{pert}}^s$ is a perturbed version of original instance $\mathbf{x}^s$ and sg represents the stop-gradient operation.

When perturbing the instance, we propose TabMix for our augmentation strategy. Given an input sample $\mathbf{x}^s$, the generator randomly masks k features and replaces their values from other instances, as shown in Figure 4.[2] Let $\odot$ denote an element-wise multiplicator and $\mathbf{m} \in \{0, 1\}^d$ a binary mask vector indicating which feature to replace. The mixing operation is defined as:

$$\text{TabMix}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{m} \odot \mathbf{x}_i + (1 - \mathbf{m}) \odot \mathbf{x}_j. \quad (12)$$

TabMix does not need to consider the scale of numeric variables; hence it is easily applicable to various datasets. The perturbed instance $\mathbf{x}_{\text{pert}}^s$ is obtained as follows:

$$\mathbf{x}' \sim \text{Sample}(\mathcal{D} \setminus \mathbf{x}^s) \quad (13)$$

$$\mathbf{x}_{\text{pert}}^s = \text{TabMix}(\mathbf{x}^s, \mathbf{x}'). \quad (14)$$

Finally, the loss function of the entire process is the sum of two losses: fairness-aware contrastive loss ($L_{\text{fair-cl}}$) and self-knowledge distillation loss ($L_{\text{self-kd}}$) as in Eq. 15.

$$L_{\text{total}} = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}^s \in \mathcal{D}} (L_{\text{fair-cl}}(\mathbf{x}^s) + L_{\text{self-kd}}(\mathbf{x}^s)) \quad (15)$$

4 EXPERIMENT

We compared DualFair with the latest models on multiple fairness-aware tabular datasets. Component analyses were performed to test the effect of each module. We also present qualitative analyses and case studies on how the model handles sensitive attributes.

4.1 Performance Evaluation

Datasets: We use the following datasets: (1) UCI Adult [3] contains 48,842 samples along with label information that indicates whether a given individual makes over 50K per year as a downstream task; (2) UCI German Credit [3] includes 1,000 samples with 20 attributes and aims to predict credit approvals; (3) COMPAS [2] includes 6,172 samples and is used to predict recidivism risk (i.e., the risk of a criminal defendant committing a crime again) of given individuals; (4) LSAC [47] contains 22,407 samples and estimates whether a given individual will pass the law school admission; (5) Students Performance (Students) [16] consists of 649 samples and predicts the grade in exam; (6) Communities [38] consists of 1,994 samples and estimates the number of violent crimes per 100K population.

---

[2]We mask 50% of the non-sensitive attributes. The masking ratio is not critical to the downstream task.

Table 1: Data descriptions of six datasets

Dataset	# samples	# attr.	Sensitive attr.	Split	Task
Adult	48,842	14	gender, race	2:1	classification
Credit	1,000	20	gender	4:1	classification
COMPAS	6,172	7	gender, race	4:1	classification
LSAC	22,407	12	gender, race	4:1	classification
Students	649	33	gender	4:1	regression
Communities	1,994	128	race	4:1	regression

We split each dataset into disjoint training and test sets. Embedding learning and counterfactual sample generation are based on the knowledge of the training set; evaluations are based on the test set. Table 1 summarizes statistics of each dataset.

Evaluation: The evaluation uses embeddings learned from the last ten epochs, and the averaged results of five metrics below:

- **AUC**, the area under the receiver operating characteristics, measures the prediction performance of the downstream classification task (i.e., the performance of the binary classifier to distinguish between cases and non-cases). If this value is 1, the model distinguishes the target variable from input instances with absolute precision.
- **RMSE**, the root mean squared error, measures the deviance between the prediction and ground-truth for the regression task.
- **Demographic Parity Distance (ΔDP)** is a group fairness metric and is defined as the expected absolute difference between the predictions of protected groups. Given a set of sensitive attributes S, the definition of ΔDP is as follows ($s \neq s'$):

$$\Delta DP = \mathbb{E}_{s,s' \in S}[|P(\hat{Y} = 1|s) - P(\hat{Y} = 1|s')|]. \quad (16)$$

- **Equalized Odds (ΔEO)**, also known as equality of opportunity, is a group fairness metric based on the expected difference between the estimated positive rates for two protected groups. Given the set of sensitive attributes S, ΔEO is defined as follows ($y \in \{0, 1\}$, $s \neq s'$):

$$\Delta EO = \mathbb{E}_{s,s' \in S}[|P(\hat{Y} = 1|s, Y = y) - P(\hat{Y} = 1|s', Y = y)|]. \quad (17)$$

- **Counterfactual Parity Distance (ΔCP)** is a metric for counterfactual fairness and measures the prediction parity between two instances in counterfactual pair relationships. Assume that $(\mathbf{x}, \hat{\mathbf{x}})$ are from a set of counterfactual pairs $\mathcal{D} \times \hat{\mathcal{D}}$. Then ΔCP is defined as follows:

$$\Delta CP = \mathbb{E}_{(\mathbf{x}, \hat{\mathbf{x}}) \in \mathcal{D} \times \hat{\mathcal{D}}}[|P(\hat{Y} = 1|\mathbf{x}) - P(\hat{Y} = 1|\hat{\mathbf{x}})|]. \quad (18)$$

Baselines: A total of nine baselines are employed. The first two directly utilize raw datasets to classify target variables, given that the downstream task is already known. They differ by input structure: (1) original data (LR) and (2) original data concatenated with the synthetic counterfactual samples (C-LR). The next one is (3) SCARF [4], an unsupervised contrastive learning method that does not consider any fairness requirement during training. The remaining six learn fair representations from data without any information on downstream tasks. They utilize different unsupervised representation learning techniques with fairness-aware objectives. (4) VFAE [34] introduces the maximum mean discrepancy term on top of the variational autoencoder to produce fair representations. (5) LAFTR [35]

adopts an adversarial approach to avoid unfair predictions from embeddings. (6,7) MIFR and L-MIFR [40] learn the controllable fair representation through mutual information. The two differ in the use of the Lagrangian dual optimization method. (8,9) C-InfoNCE and WeaC-InfoNCE [43] maximize the conditional mutual information within representations. The two differ in how they introduce the sensitive variable in the InfoNCE objective. For all methods, we use a logistic regression model for the downstream classification task and a Random Forest regression model for the downstream regression task as a base predictor. The predictor is learned on top of either raw dataset (1–2) or learned embeddings (3–9).

Implementation details: We trained the multi-layer perceptron (MLP) model. A three-layer MLP was used for the backbone network f, and a two-layer MLP for both projection and prediction heads g, (h_1, h_2). ReLU activation function was used for all architectures, with the training of 200 epochs and a batch size of 128. Adam optimizer with a learning rate of 1e-3 and a weight decay factor of 1e-6 was utilized. For the counterfactual sample generator, three-layer, two-layer, and two-layer MLPs were utilized for the encoder, decoder, and discriminator, respectively. The sample generator was trained for 600 epochs with the Adam optimizer. Mode-specific normalization [49] was used for preprocessing categorical variables.

Table 2: Performance comparison summaries among fairness-aware baselines and DualFair. Averaged rank for each evaluation metric across six datasets is reported.

Method	AUC/RMSE	ΔDP	ΔEO	ΔCP	Total
VFAE	3.2	5.8	5.0	4.2	4.6
LAFTR	**2.7**	6.2	5.5	2.8	4.3
MIFR	6.0	3.0	3.8	4.6	4.4
L-MIFR	5.3	3.2	3.3	3.8	3.9
C-InfoNCE	3.2	4.5	3.3	5.2	4.1
WeaC-InfoNCE	4.7	2.8	4.3	4.8	4.2
DualFair	3.0	**2.5**	**2.5**	**2.2**	**2.6**

Results: DualFair outperforms other baselines in terms of the averaged rank for each evaluation metric. Table 2 shows that our model's embedding maintains its prediction performance (i.e., AUC and RMSE) while successfully removing any bias related to sensitive information from the embedding.

Tables 3 and 4 report the detailed performance of DualFair and other baselines on downstream classification and regression tasks. According to the results, naively adding the counterfactual samples (i.e., C-LR) is insufficient to handle bias from sensitive information. Similarly, the fair embedding learning baselines such as VFAE, L-MIFR, or WeaC-InfoNCE fail to remove sensitive information entirely nor generate a fair embedding without losing critical information for the downstream task. These findings suggest that DualFair achieves both the performance and fairness requirements in a single training.

4.2 Component Analyses

We performed an ablation study by repeatedly assessing and comparing the models after removing each component. We also examined the counterfactual sample quality.

Table 3: Detailed results over downstream classification tasks. Across the four datasets, *gender* is considered a sensitive attribute to protect. Up-arrow (↑) denotes that a higher value is better, and down-arrow (↓) is vice versa. The best results among fair representation learning baselines are highlighted in bold.

Method	UCI Adult				UCI German credit				COMPAS				LSAC			
	AUC (↑)	ΔDP (↓)	ΔEO (↓)	ΔCP (↓)	AUC	ΔDP	ΔEO	ΔCP	AUC	ΔDP	ΔEO	ΔCP	AUC	ΔDP	ΔEO	ΔCP
LR	0.9055	0.1931	0.1763	0.0812	0.7580	0.0653	0.1066	0.1765	0.7354	0.1175	0.1733	0.0873	0.8515	0.0463	0.0378	0.0493
C-LR	0.9023	0.1610	0.1679	0.0568	0.7252	0.0161	0.0879	0.1416	0.7321	0.0441	0.0592	0.0286	0.8513	0.0621	0.0361	0.0327
SCARF	0.9012	0.1905	0.1929	0.1097	0.7459	0.0719	0.1443	0.1460	0.7384	0.1155	0.1747	0.1029	0.8544	0.0716	0.0357	0.0477
VFAE	0.8249	0.1764	0.1995	0.1126	0.6582	0.0365	0.0514	0.2785	0.7089	0.0286	0.1069	0.1226	**0.8275**	0.0725	0.0293	0.0569
LAFTR	**0.8943**	0.1727	0.1720	0.0746	0.6845	0.0223	0.0282	0.1917	0.7142	0.0486	0.1055	0.0562	0.6380	0.0731	0.0172	0.0428
MIFR	0.8162	0.0427	0.0649	0.1238	0.5463	0.0113	0.0297	0.1241	0.6112	**0.0081**	0.0423	0.1049	0.7672	0.0676	**0.0142**	0.0618
L-MIFR	0.8211	0.0337	0.0578	0.1408	0.5703	**0.0045**	0.0143	0.1217	0.6243	0.0100	**0.0409**	0.0996	0.7694	0.0680	0.0152	0.0613
C-InfoNCE	0.8252	0.0884	0.0909	0.1492	0.6719	0.0206	0.0191	0.1141	0.7192	0.0495	0.0859	0.0793	0.7977	0.0743	0.0149	0.2647
WeaC-InfoNCE	0.8153	0.0786	0.0802	0.1895	0.6872	0.0164	0.0165	**0.1023**	0.7130	0.0394	0.0871	0.0750	0.7914	0.0648	0.0412	0.2048
DualFair	0.8002	**0.0285**	**0.0435**	**0.0644**	0.7292	0.0061	**0.0130**	0.1454	0.7235	0.0401	0.0622	**0.0409**	0.7790	**0.0527**	0.0239	**0.0361**

Table 4: Detailed results over downstream regression tasks. The best results among fair representation learning baselines are highlighted in bold.

Method	Students			Communities		
	RMSE (↓)	ΔDP (↓)	ΔCP (↓)	RMSE	ΔDP	ΔCP
LR	0.0392	0.0177	0.0794	0.0145	0.2146	0.1559
C-LR	0.0399	0.0081	0.0777	0.0141	0.1643	0.0854
SCARF	0.0746	0.0196	0.0412	0.0821	0.1631	0.0881
VFAE	0.0425	0.0084	**0.0519**	**0.0191**	0.1714	0.0844
LAFTR	0.0371	0.0223	0.0758	0.0194	0.1725	0.1311
MIFR	0.0475	0.0056	0.0785	0.0261	0.1139	0.0916
L-MIFR	0.0485	0.0063	0.0797	0.0258	0.1161	0.0902
C-InfoNCE	0.0415	**0.0049**	0.0626	0.0317	0.0841	0.0884
WeaC-InfoNCE	0.0407	0.0054	0.0676	0.0319	**0.0813**	0.0891
DualFair	**0.0382**	0.0054	0.0735	0.0247	0.1130	**0.0816**

Table 5: Ablation results of DualFair on UCI Adult dataset. L_{align} and $L_{\text{distribution}}$ are from fairness-aware contrastive loss while $L_{\text{self-kd}}$ is self-knowledge distillation loss.

Setup	AUC	ΔDP	ΔEO	ΔCP
DualFair	0.80	0.03	0.04	0.06
w/o $L_{\text{self-kd}}$	0.77	0.02	0.03	0.05
w/o L_{align}	0.78	0.04	0.04	0.08
w/o L_{align} and $L_{\text{distribution}}$	0.81	0.12	0.12	0.08

Table 6: The model trained with counterfactual samples shows good classification performance on the original data, proving that counterfactual samples are augmented well.

Training set	Adult		Credit		Compas		LSAC	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Original	0.91	0.78	0.76	0.68	0.74	0.68	0.85	0.65
Counterfactual	0.89	0.76	0.75	0.62	0.73	0.67	0.84	0.63

Ablation study: DualFair utilizes two learning objectives: fairness-aware contrastive loss to ensure both group and counterfactual fairness, and self-knowledge distillation loss to ensure the representation quality. Ablations remove each loss objective from the full model to assess the unique contribution. Table 5 reports the results based on the UCI Adult dataset. The full model achieves the best balance between fairness and prediction performance, implying that each loss plays a unique role in designing fair representations. The result without self-knowledge distillation shows that our fairness-aware contrastive loss effectively improves fairness while there is a trade-off for prediction performance. On the other hand, the experiment only with self-knowledge distillation loss (i.e., w/o L_{align} and $L_{\text{distribution}}$) produces opposite result. The result without only alignment loss, which is in charge of counterfactual fairness, contrasts the role of two losses in fairness-aware contrastive objective; Counterfactual fairness does not have any improvement while group fairness has been reduced. Additionally, we experimented using Gaussian noise and dropout as alternative augmentation strategies for TabMix. This ablation reduced the prediction accuracy (AUC: 0.80→0.78, 0.75) while maintaining fairness (DP: 0.03→0.03, 0.02).

Counterfactual samples: To test the quality of the generated counterfactual samples, we examined if the target variable is predictable, even if the model is trained only with the synthetic counterfactual samples. Table 6 reports the logistic regression performance that changes the training set from the original data. The latter model is on par with the model trained with the original dataset.

We next examined if feature correlations are maintained for counterfactual samples. Figure 5 shows the correlation matrix between features in the original UCI adult dataset and the counterfactual dataset. Pearson correlation is used between continuous variables, and Cramer's V value is utilized between categorical variables. To measure the correlation between categorical and continuous variables, we label-encode the categorical variables to their continuous counterparts and compute the Pearson correlation with original continuous values. Two matrices show a remarkable resemblance,

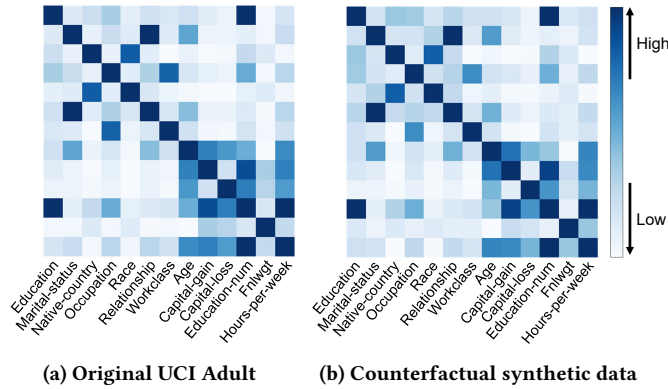**(a) Original UCI Adult** **(b) Counterfactual synthetic data**

Figure 5: Correlation heatmaps show that (a) original UCI adult dataset and (b) synthetic dataset from our counterfactual sample generator—follow similar patterns.
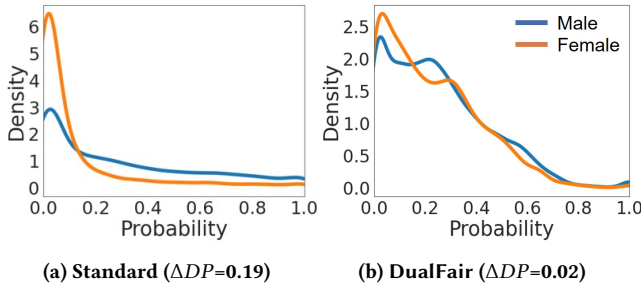

(a) Standard (ΔDP=0.19) **(b) DualFair (ΔDP=0.02)**

Figure 6: Income prediction results between the male and female groups are similar when we use DualFair's embeddings. We compared two models, (a) standard logistic regression on top of original data and (b) logistic regression on top of DualFair embedding, on the UCI Adult dataset.

indicating that the relationship between features in the counterfactual samples is well-maintained.

4.3 Qualitative Analysis

We visually check the learned representations to investigate how well our model handles sensitive information. Figure 6 shows how debiasing achieves group fairness in downstream predictions. It compares the prediction results on UCI Adult for two models: standard logistic regression applied on the raw dataset and the same regression model using the DualFair embeddings. After debiasing, there is almost no difference in the distribution of predicted income between male and female groups, as depicted in Figure 6b. However, two probability density functions in Figure 6a appear substantially different in the standard model. These results confirm the outstanding debiasing potential of the proposed model.

We also investigate how well DualFair achieves counterfactual fairness by examining the difference in the ΔCP values of the proposed model and its ablation GroupFair, which omits the alignment loss between counterfactual pairs in fairness-aware contrastive loss (Eq. 10). Experimental comparison over UCI Adult in Figure 7a shows a smaller prediction difference for the full model, indicating
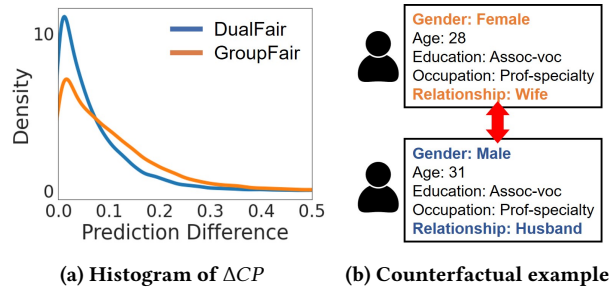

(a) Histogram of ΔCP **(b) Counterfactual example**

Figure 7: Additional findings on the counterfactual model: (a) Comparison of the ΔCP value between DualFair and its ablation GroupFair which omits the alignment loss between counterfactual pairs. (b) Observable variables affected by sensitive attributes are also changed in the counterfactual example.

that the omitted loss is effective in debiasing sensitive attributes at the individual level. One of the counterfactual pairs from the model is illustrated in Figure 7b. Note that in generating counterfactual examples, we do not simply flip sensitive attributes (i.e., gender) but also observable variables (i.e., age, relationship) change together. When we compute ΔCP for the example case from two models, we confirm that the proposed DualFair satisfies the fairness concerns to some extent (prediction probability for original: 0.22 vs. counterparts: 0.24). Meanwhile, GroupFair fails to debias gender information and gives a higher score to the male counterparts (0.24 vs. 0.45).

5 CONCLUSION

We presented DualFair, a self-supervised embedding learning model that de-biases sensitive data attributes without any prior information on downstream tasks. Its design includes a unique fairness-aware contrastive loss and self-knowledge distillation technique. Experiments confirm that DualFair generates rich data representations that ensure both group fairness and counterfactual fairness. Our model is applicable to various Web applications, including classification, ranking, recommendation, and text generation tasks.

The algorithmic bias observed in search engine results and social media platforms has reinforced the need for a clear policy for protecting sensitive attributes. However, the problem is more complex because bias can also exist in other domains, such as in natural language processing (e.g., Q&A generation, chatbot). The recently released 'AI Bill of Rights' blueprint from the White House states that "you should not face discrimination by algorithms and systems should be used and designed in an equitable way." We believe that our self-supervised learning with debiasing techniques can serve as a building block for advancing the performance and fairness requirements of real-world web applications.

REFERENCES

[1] Mingxiao An, Fangzhao Wu, Chuhan Wu, Kun Zhang, Zheng Liu, and Xing Xie. 2019. Neural news recommendation with long-and short-term user representations. In *proc. of the ACL*. 336–345.

[2] Julia Angwin, Kirchner Larson, Jeff Kirchner, and Surya Mattu. 2016. Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica* (2016).

[3] Arthur Asuncion and David Newman. 2007. UCI machine learning repository.

[4] Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. 2022. SCARF: Self-Supervised Contrastive Learning using Random Feature Corruption. In *proc. of the ICLR*.

[5] Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial Removal of Demographic Attributes Revisited. In *proc. of the EMNLP-IJCNLP*. 6329–6334.

[6] Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *proc. of the ACM FAT**. 514–524.

[7] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*. 4349–4357.

[8] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. 2015. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 51, 1 (2015), 22–45.

[9] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *proc. of the FAT**. 77–91.

[10] Robin Burke. 2017. Multisided fairness for recommendation. *arXiv preprint arXiv:1707.00093* (2017).

[11] Robin Burke, Nasim Sonboli, and Aldo Ordonez-Gauger. 2018. Balanced neighborhoods for multi-sided fairness in recommendation. In *proc. of the FAT**. 202–214.

[12] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *proc. of the FAccT*. 339–348.

[13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *proc. of the ICML*. 1597–1607.

[14] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *proc. of the CVPR*. 15750–15758.

[15] Vincent Conitzer, Rupert Freeman, Nisarg Shah, and Jennifer Wortman Vaughan. 2019. Group fairness for the allocation of indivisible goods. In *proc. of the AAAI*, Vol. 33. 1853–1860.

[16] Paulo Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance. (2008).

[17] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *proc. of the ITCS*. 214–226.

[18] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *proc. of the ITCS*. 214–226.

[19] Yanai Elazar and Yoav Goldberg. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *proc. of the EMNLP*. 11–21.

[20] Pratik Gajane and Mykola Pechenizkiy. 2017. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184* (2017).

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in Neural Information Processing Systems* 33 (2020), 21271–21284.

[22] Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse Adversaries for Mitigating Bias in Training. In *proc. of the EACL*. 2760–2765.

[23] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *proc. of the NeurIPS*. 3315–3323.

[24] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *proc. of the CVPR*. 9729–9738.

[25] Mitchell Hoffman, Lisa B Kahn, and Danielle Li. 2018. Discretion in hiring. *The Quarterly Journal of Economics* 133, 2 (2018), 765–800.

[26] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *proc. of the ECML PKDD*. 35–50.

[27] Kyungyul Kim, ByeongMoon Ji, Doyoung Yoon, and Sangheum Hwang. 2021. Self-knowledge distillation with progressive refinement of targets. In *proc. of the ICCV*. 6567–6576.

[28] Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *proc. of the ICLR*.

[29] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *proc. of the Seventh Joint Conference on Lexical and Computational Semantics*. 43–53.

[30] Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. 2019. Generalized sliced wasserstein distances. *Advances in Neural Information Processing Systems* 32 (2019).

[31] Öykü Deniz Köse and Yanning Shen. 2021. Fairness-Aware Node Representation Learning. *arXiv preprint arXiv:2106.05391* (2021).

[32] Matt Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *proc. of the NeurIPS*. 4069–4079.

[33] Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards Robust and Privacy-preserving Text Representations. In *proc. of the ACL*. 25–30.

[34] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. 2016. The Variational Fair Autoencoder. In *proc. of the ICLR*.

[35] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *proc. of the ICML*. 3384–3393.

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).

[37] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *proc. of the ACL*. 7237–7256.

[38] Michael Redmond and Alok Baveja. 2002. A data-driven software tool for enabling cooperative information sharing among police departments. *European Journal of Operational Research* 141, 3 (2002), 660–678.

[39] Chris Russell, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. 2017. When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness. In *Advances in Neural Information Processing Systems*. 6414–6423.

[40] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2019. Learning controllable fair representations. In *proc. of the AIStat*. 2164–2173.

[41] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.

[42] Ajinkya Tejankar, Soroush Abbasi Koohpayegani, Vipin Pillai, Paolo Favaro, and Hamed Pirsiavash. 2021. ISD: Self-supervised learning by iterative similarity distillation. In *proc. of the ICCV*. 9609–9618.

[43] Yao-Hung Hubert Tsai, Martin Q Ma, Han Zhao, Kun Zhang, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Conditional Contrastive Learning: Removing Undesirable Information in Self-Supervised Representations. *arXiv preprint arXiv:2106.02866* (2021).

[44] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *IEEE/ACM International Workshop on Software Fairness*. 1–7.

[45] Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *proc. of the ICML*. 9929–9939.

[46] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In *proc. of the ICCV*. 5309–5318.

[47] Linda F Wightman. 1998. LSAC National Longitudinal Bar Passage Study. LSAC Research Report Series. (1998).

[48] Chuhan Wu, Fangzhao Wu, Xiting Wang, Yongfeng Huang, and Xing Xie. 2021. Fairness-aware News Recommendation with Decomposed Adversarial Learning. In *proc. of the AAAI*. 4462–4469.

[49] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling Tabular data using Conditional GAN. *Advances in Neural Information Processing Systems* 32 (2019), 7335–7345.

[50] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2017. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. In *proc. of the WWW*. 1171–1180.

[51] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. 2019. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research* 20, 1 (2019), 2737–2778.

[52] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *proc. of the AIES*. 335–340.

Table 7: Additional results over downstream classification tasks in which *race* is set as a target-sensitive attribute to protect. Performances are reported with four evaluation metrics (AUROC, ΔDP, ΔEO, and ΔCP) over three datasets.

Method	UCI Adult				COMPAS				LSAC			
	AUROC ($\uparrow$)	ΔDP ($\downarrow$)	ΔEO ($\downarrow$)	ΔCP ($\downarrow$)	AUROC	ΔDP	ΔEO	ΔCP	AUROC	ΔDP	ΔEO	ΔCP
LR	0.9055	0.0867	0.1028	0.0771	0.7354	0.1214	0.2515	0.0608	0.8515	0.1306	0.1081	0.0448
C-LR	0.9046	0.0729	0.0854	0.0490	0.7353	0.1050	0.3222	0.0273	0.8504	0.1499	0.1462	0.0293
SCARF	0.9011	0.0792	0.0928	0.1058	0.7418	0.1507	0.2949	0.0802	0.8515	0.1399	0.1351	0.0523
VFAE	0.8445	0.0945	0.0586	0.0746	0.7039	0.1319	0.1664	0.3115	0.8218	0.0595	0.1472	0.1314
LAFTR	0.8940	0.0652	0.0667	0.0878	0.7175	0.0378	0.1373	0.3257	0.7213	0.0578	0.1031	0.0129
MIFR	0.8297	0.0931	0.0255	0.0463	0.6454	0.1035	0.0099	0.2104	0.7129	0.0579	0.0181	0.0084
L-MIFR	0.8392	0.1044	0.0250	0.0455	0.6895	0.1077	0.0593	0.1819	0.7188	0.0566	0.0277	0.0082
C-InfoNCE	0.8758	0.1854	0.0673	0.0716	0.6256	0.0462	0.0327	0.2091	0.6548	0.0647	0.0141	0.0052
WeaC-InfoNCE	0.8843	0.1911	0.0687	0.0895	0.6084	0.0325	0.0313	0.0898	0.7087	0.1015	0.0267	0.0118
DualFair	0.8067	0.0527	0.0205	0.0355	0.7215	0.0376	0.1096	0.3112	0.7553	0.0491	0.0489	0.0451

A APPENDIX

A.1 Implementation Details

DualFair: DualFair exploits the multi-layer perceptron architecture for both the main model and counterfactual sample generator. The main model consists of the backbone network (f), the projection head (g), and two prediction heads (h_1, h_2). We chose all components to be ReLU networks, whose hidden dimension is set to 256. The backbone network has three layers, and all heads have two layers. DualFair is trained 200 epochs, and its batch size is set to 128. The Adam optimizer with a learning rate of 1e-3 and a weight decay factor of 1e-6 is adopted. We apply one-hot encoding to discrete variables for input preprocessing and z-score scaling to continuous variables.

Counterfactual sample generator: We introduce the C-VAE model to generate counterfactual samples. When preprocessing the input data, discrete values are encoded into the one-hot vector, and every continuous value is converted to a vector of probability density via mode-specific normalization [49]. Mode-specific normalization uses a variational Gaussian mixture model (VGM) to estimate the number of modes and fit the Gaussian mixture model on top of the target distribution. Then, the probability density of each mode is computed for the normalization. We discovered that mode-specific normalization improves counterfactual sample quality more than naive min-max normalization. The counterfactual sample generator is trained for 600 epochs, and the batch size is set to 256. The weight between reconstruction loss and probability distribution loss is set to 2:1.

Computational complexity: We used four A100 GPUs for all experiments. Our model took 20% more training time than WeaC-InfoNCE (11 vs. 9 minutes for 200 epochs), and the counterfactual VAE training took less than 5 minutes.

A.2 Further Results on Performance Evaluation

We present the performance of DualFair and other baselines on downstream classification tasks by setting gender as the sensitive attribute in the main manuscript. We here present the extra results on race to support DualFair's generalizability on different sensitive attributes. In the race attribute, UCI Adult, COMPAS, and LSAC have five, five, and six classes, respectively. UCI German Credit dataset does not include race information and hence is skipped.

The evaluation results in Table 7 demonstrate that DualFair learns data distribution of critical features while minimizing spurious information from the multi-class sensitive attribute.

A.3 Further Results on Ablation study

To confirm the effectiveness of the proposed counterfactual sample generator, we further assessed two ablations on loss objectives: (1) without cyclic consistency loss L_{cyc} (Eq. 3) and (2) without reconstruction loss in L_{vae} (Eq. 1). Table 8 reports the test set performance of the logistic regression model, after fitting with synthetic counterfactual samples from each baseline. When we conduct experiments over four datasets, our model with full components outperforms baselines in terms of both the AUC and F1-score.

Table 8: Ablation study of counterfactual sample generator. The proposed model with full components shows the best classification performance on the original data.

Training set	Adult		Credit		Compas		LSAC	
	AUC	F1	AUC	F1	AUC	F1	AUC	F1
Original	0.91	0.78	0.76	0.68	0.74	0.68	0.85	0.65
Ours	0.89	0.76	0.75	0.62	0.73	0.67	0.84	0.63
w/o Cyclic loss (L_{cyc})	0.88	0.73	0.71	0.61	0.72	0.67	0.84	0.66
w/o Reconstruction in (L_{vae})	0.35	0.45	0.35	0.45	0.48	0.42	0.36	0.47

A.4 Further Results on Qualitative Analysis

Figure 8 visualizes the embeddings over the UCI Adult dataset from four models: SCARF and DualFair at three different points of training (1, 10, and 200 epoch). SCARF does not consider any fairness requirements, and hence the learned representations of the same protected group (i.e., gender) are placed nearby, forming a local cluster in the embedding space (see Fig. 8a). When we fit the logistic regression model to predict the sensitive attribute on top of the embeddings, the model reports very high AUC values. Meanwhile, the sensitive information is debiased across the epochs, and the sensitive attribute becomes indistinguishable in the embedding space (see Fig. 8b–8d). The AUC result in this new embedding gradually decreases with the training epoch.

(a) SCARF
(AUC=0.92)

(b) DualFair 1 epoch
(AUC=0.75)

(c) DualFair 10 epochs
(AUC=0.72)
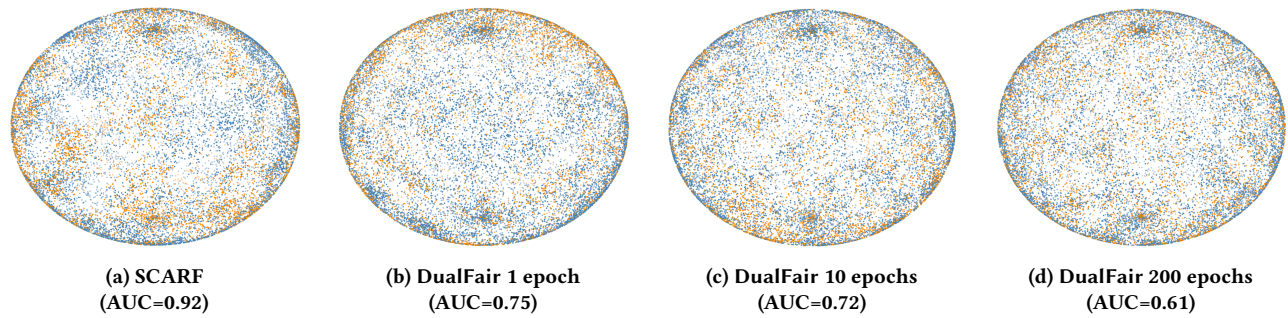
(d) DualFair 200 epochs
(AUC=0.61)

Figure 8: UMAP visualization of trained embeddings over UCI Adult dataset from (a) SCARF and (b-d) DualFair with different epochs. Each dot represents the embedding of each individual, and color displays his/her gender. With these embeddings, we fit a classifier to predict the sensitive attribute (i.e., gender) of the test instances, anticipating that sensitive attribute cannot be determined from DualFair embeddings. As we expected, classifying the sensitive attribute using DualFair embeddings was difficult (i.e., AUC=0.61). We can infer that sensitive information was avoided in the process of creating the embedding.