# FADES: Fair Disentanglement with Sensitive Relevance

Taeuk Jang, Xiaoqian Wang*
Purdue University
465 Northwestern Ave, West Lafayette, IN 47907, USA
{jang141@,joywang}@purdue.edu

## Abstract

*Learning fair representation in deep learning is essential to mitigate discriminatory outcomes and enhance trustworthiness. However, previous research has been commonly established on inappropriate assumptions prone to unrealistic counterfactuals and performance degradation. Although some proposed alternative approaches, such as employing correlation-aware causal graphs or proxies for mutual information, these methods are less practical and not applicable in general. In this work, we propose FAir DisEntanglement with Sensitive relevance (FADES), a novel approach that leverages conditional mutual information from the information theory perspective to address these challenges. We employ sensitive relevant code to direct correlated information between target labels and sensitive attributes by imposing conditional independence, allowing better separation of the features of interest in the latent space. Utilizing an intuitive disentangling approach, FADES consistently achieves superior performance and fairness both quantitatively and qualitatively with its straightforward structure. Specifically, the proposed method outperforms existing works in downstream classification and counterfactual generations on various benchmarks.*

## 1. Introduction

Deep generative models have made great accomplishments in various applications including style transfer [54], video generation [5], image translation [22], text-to-image generation [9], etc. Among them, variational autoencoder (VAE) based models [21, 26, 38] significantly contributed to the evolution of deep learning and generative models by offering a better understanding of the latent space and stable performance. Although VAEs have been widely adopted due to the compactness and robustness of the learned representation, there is plenty of room for improvement in various aspects. One of the highlighted areas is fairness since preventing discrimination is crucial to gaining trustworthiness of deep learning models in practice.

While deep models achieve superior performance by exploiting high dimensional latent representation, the deep representations are susceptible to absorbing spurious correlations with sensitive information [29, 30, 41], which can lead to potential fairness violations and biased outcomes. To address the problem, a line of work was proposed to learn fair representations. Some [29, 46] proposed to learn latent representation that is invariant to sensitive information, while others [6, 10] aimed to disentangle the latent features into two subsets: target (non-sensitive) and sensitive codes. This is often done by introducing regularization terms for fair representation learning in addition to performance objectives. Specifically, fair disentanglement approaches minimize mutual information between the subsets [3] to separate the target label and sensitive information.

However, we argue that achieving such fair disentanglement is unattainable under data bias, which is one of the prevalent causes of fairness violations. Many datasets [11, 27] contain discriminatory labels that are influenced by societal biases and stereotypes, which are unfavorable to certain demographics [39]. This leads to an unwanted correlation between the target and sensitive attributes in practice. When this correlation is overlooked, the fairness and performance goals of the disentanglement methods contradict each other and cannot be achieved simultaneously. Specifically, the fairness goal aims at learning target and sensitive codes to be independent, while the performance goal pursues perfectly recovering the target and sensitive information from the respective codes. However, when the target label and sensitive attribute are inherently correlated, it is impossible to achieve both goals at the same time.

Moreover, some features may contain both sensitive and target information, making it difficult to protect sensitive information while making accurate predictions in practice [13]. This makes the objective of previous works, which aimed to learn perfectly separable subsets, unattainable. For instance, in the hair color estimation task from portraits

*Corresponding author. This work was partially supported by NSF IIS #1955890, IIS #2146091.

[58], which is known to be gender-biased, the *bald* attribute is related to both *hair color* and *gender*. Therefore, such fair disentanglement can be achieved only under the naive assumption: target and sensitive attributes are independent, where no attribute contains information related to both target and sensitive information.

Some [25, 57] attempted to address this by introducing complex causal models by considering more complicated relations between features. However, these require domain knowledge to understand the causality between the features and build a comprehensive graph, which is demanding and not always accessible. Without accurately knowing the causal relationships among the attributes, the quality of the generated counterfactuals is significantly impaired, *e.g.,* a girl with a mustache.

In this work, we propose a novel approach in disentanglement learning for fairness to address the limitations by introducing *sensitive relevant* code. Our method effectively directs the information correlated to both target and sensitive information to sensitive relevant code by ensuring conditional independence without requiring domain knowledge. To achieve this, we minimize conditional mutual information as it sets an upper bound for conditional independence, which encourages learning genuinely fair disentanglement, *i.e.,* independence between target and sensitive codes. As a result, the proposed method provides fair representation with improved control and interpretation while outperforming existing works in both fairness and utility.

The contribution of the work can be summarized as follows:

1. We theoretically demonstrate that the fairness and performance goals of previous works inherently contradict under common data bias, resulting in unstable training and degraded utility.
2. We propose a novel approach to address the contradiction by introducing the sensitive relevant code, which theoretically leads to optimal disentanglement of sensitive and target codes.
3. We propose a framework to disentangle features in the information theory perspective by leveraging conditional mutual information. The proposed method has a simple structure but is effective and does not require domain knowledge.
4. We empirically validate the effectiveness of the proposed method on multiple benchmark datasets.

## 2. Related Work

### 2.1. Learning Fair Representation

The goal of learning fair representation is to make accurate predictions on downstream tasks while filtering out the influence of sensitive information. Simply removing sensitive information, *i.e.,* fairness through blindness, is not sufficient to achieve fairness as there exist related features from which we can infer sensitive information. In classification, Madras *et al.* [36] proposed to constrain group fairness metrics, *e.g.,* equalized odds [18] and demographic parity, as the adversarial objectives. Besides, the follow-up method integrates the whole procedure of optimization in a single neural network to improve stability of the adversarial network [1]. Also, FairGAN [52] focuses on generating fair data aiming to fool a strong discriminator to recognize which sensitive group a generated image belongs to. However, such discrimination-free representation requires strong conditions and has degradation in performance in image synthesis since sensitive information is also essential for image generation.

### 2.2. Disentanglement Learning

Instead of learning sensitive-information-free representation, a line of works proposed to disentangle the representation into two sets of latent variables: target and sensitive code. $\beta$-VAE [21] initiated disentangling semantic features by encouraging the variational distribution to satisfy stronger prior constraint. FactorVAE [24] proposed to minimize the total correlation for further decomposition and its variant [32] further obfuscate sensitive information in the latent space. FFVAE [6] applied the total correlation [3] on disentangling sensitive code flexible to the dimension of sensitive information. GVAE [10] also employed adversaries to ensure minimizing the leakage of unwanted information in each latent code. ODVAE [47] proposed to learn target and sensitive code to follow orthogonal priors while FairDisCo [31] minimized distance covariance, serving as a non-adversarial alternative to enforce independence.

However, we claim that strictly disentangling the original space into two perfectly independent subsets of latent codes is unachievable. The primary objective of representation learning is to recover sensitive and target information from each latent code. However, existing fair methods mostly overlooked that the target and sensitive information is often correlated in practice, which is the major source of the data bias [39]. Therefore, if sensitive and target codes can perfectly recover sensitive and target information, respectively, they inherently cannot be independent in nature. For instance, facial attribute recognition task on CelebA dataset [34] shows that attributes such as *mustache* (sensitive relevant) is both related to *gender* (sensitive) and *attractiveness* (target).

### 2.3. Counterfactual Fairness

The objective of counterfactual fairness [28] is to constrain biased decisions when the protected attributes are perturbed at the individual level, which is evaluated by the potential outcomes of altering sensitive attributes of individuals while keeping other features the same, *i.e.,* counterfactuals. The

causal inference technique is utilized to generate the counterfactual instances [4, 51, 57]. To comprehend the complex connections between features, some studied the causal effect between features by building a graph-based model with predefined intervention and individual variables [25]. However, without accurately knowing the causal relationships among the attributes, *i.e.,* domain knowledge, the quality of the generated counterfactuals can be significantly impaired and negatively affect the utility. For instance, lacking domain knowledge can generate unrealistic data, such as a boy attending a girl's high school.

Unlike existing works, we introduce *sensitive relevant* code to learn fair representation. We impose independence between target and sensitive information conditioned on the sensitive relevant code by minimizing conditional mutual information (CMI), which results in the independence between sensitive and target codes. Moyer *et al.* [40] utilize mutual information to learn sensitive-invariant representation. However, learning a representation that is independent of sensitive information is too restrictive [47]. While a recent work applied CMI in RL [12], to the best of our knowledge, the proposed method is the first to learn a fair representation that accounts for data bias by employing CMI.

## 3. Preliminaries

In this section, we discuss different approaches to achieving fair representation. Many works utilize variational inference following VAE [26] that approximate inference by maximizing evidence lower bound (ELBO):

$$\mathcal{L}_{\text{ELBO}}(\theta,\phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})),$$

where input $\mathbf{x}$ is fed to encoder $f(\mathbf{x};\phi)$ and decoder $g(\mathbf{z};\theta)$ reconstructs the input $\mathbf{x}$ from the given latent posterior $\mathbf{z}$. The prior $p(\mathbf{z})$ is usually set to isotropic Gaussian distribution. With $\beta > 1$, the encoder is encouraged to have a stronger agreement with the prior, which leads to factorized disentanglement [21].

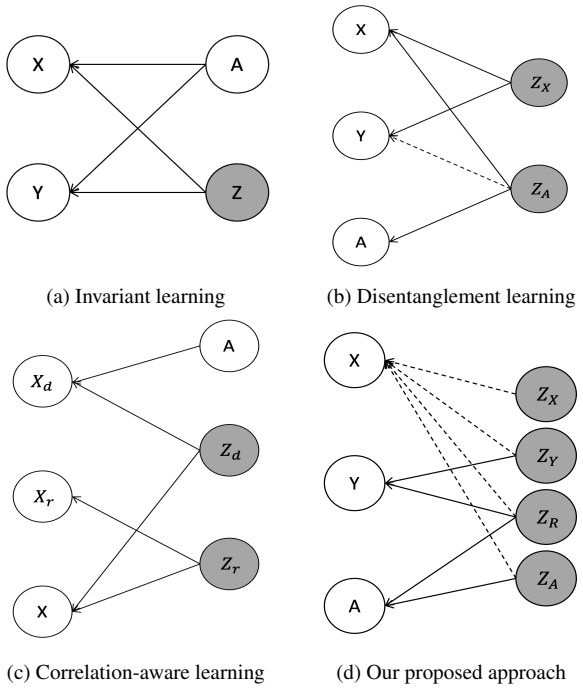
FactorVAE [24] imposes independence in latent dimensions by minimizing total correlation for disentanglement:

$$\mathcal{L}_{\text{TC}} = KL\big(q(\mathbf{z})||\prod_j \mathbf{z}_j\big) \approx \mathbb{E}_{q_\phi(\mathbf{z})}\left[\log \frac{D(\mathbf{z})}{1-D(\mathbf{z})}\right], \quad (1)$$

where discriminator $D$ seeks whether the latent code is sampled from aggregate posterior or a combination of in-batch permutations of marginal distributions across each latent dimension $j$ while the encoder tries to fool the discriminator. Figure 1 depicts the graphical models of each approach for fair representation learning.

### 3.1. Invariant Learning

Invariant learning aims to learn a unified latent variable $\mathbf{z}$ that is invariant to sensitive attribute $A$ as in Figure 1a. The




(a) Invariant learning (b) Disentanglement learning

(c) Correlation-aware learning (d) Our proposed approach

Figure 1. Graphical models of fair representation learning methods. Dashed line indicates the connection may or may not exist, depending on different model designs.

data $X$ and label $Y$ are causal downstream of sensitive attribute $A$ and unobserved variable $\mathbf{z}$. The goal is to achieve $\mathbf{z}$ independent of $A$ so that the predicted $Y$ is subsequently invariant to sensitive information. Adversarial learning [32, 35, 37] is often employed to fool a discriminator which tries to predict sensitive information from $\mathbf{z}$. While others [15, 31, 44] employ a regularizer to minimize some divergence $D$ (*e.g.,* KL divergence, MMD [16], distance covariance) among the aggregate posteriors $q_\phi(\mathbf{z}|A=a_k)$.

### 3.2. Disentanglement Learning

This group of studies disentangles the latent variable into the target and sensitive codes. Since the independence condition between $\mathbf{z}$ and $A$ is too restrictive, some works hypothesize decomposing the latent codes into two sets would promote no leakage of sensitive information from $\mathbf{z}_X$ by disentangling the sensitive information to $\mathbf{z}_A$ as in Fig 1b.

Creager *et al.* [6] proposed to minimize the mutual information between $\mathbf{z}_X$ and $\mathbf{z}_A$ leveraging total correlation [3, 24]. Following the previous works, a binary adversary learns to distinguish aggregate posterior $q_\phi(\mathbf{z}_X, \mathbf{z}_A)$ from a product of the marginals $q_\phi(\mathbf{z}_X)\prod_j q_\phi(\mathbf{z}_A^{(j)})$ where superscript $j$ denotes factorized subspace. This is empirically approximated by randomly permuting sensitive latent code $z_A$ within a batch.

Given the variables of interest $T = \{Y, A\}$, GVAE [10] incentivizes certain attribute $t \in T$ to be projected to spe-

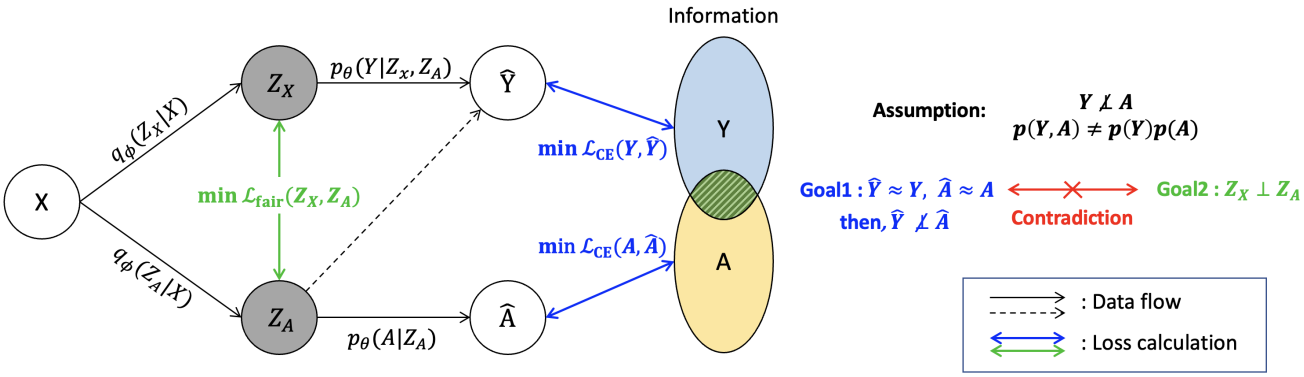



Figure 2. Illustration of the contradiction between the goals imposed by fairness loss $\mathcal{L}_{\text{fair}}$ and performance loss $\mathcal{L}_{\text{CE}}$ of disentanglement learning. Two goals conflict since target label $Y$ and sensitive attribute $A$ are not independent by data bias (green-shaded region) in practice.

cific subspaces $\mathbf{z}_t \in \mathbf{z}$ and have no leakage of the information in the remaining code $\mathbf{z}_t^{rst} = \mathbf{z} \backslash \mathbf{z}_t$. They employ two sets of classifiers: $\{w_t\}$ and $\{\xi_t\}$, where $\{w_t\}$ conveys target features, while $\{\xi_t\}$ is for preventing information leakage, respectively, for promoting disentanglement.

ODVAE [47] points out the potential convergence problem of the adversarial approaches [40, 42]. They relaxed the independence condition by encouraging $\mathbf{z}_X$ and $\mathbf{z}_A$ to follow orthogonal priors: $p(\mathbf{z}_X) = \mathcal{N}(\boldsymbol{\mu}_X, I)$ and $p(\mathbf{z}_A) = \mathcal{N}(\boldsymbol{\mu}_A, I)$, where $\boldsymbol{\mu}_X^\top \boldsymbol{\mu}_A = 0$. However, it requires hard-coded orthogonal priors and the orthogonality does not guarantee independence.

### 3.3. Correlation-Aware Learning

It is suggested that employing a more sophisticated causal graph may facilitate fair disentanglement, demonstrating improved counterfactual generations. These models separate the latent variables based on their correlation with sensitive information, as in Figure 1c. For instance, in the case of *gender* as $A$, they define the descendant attributes of $A$ as $X_d = \{\text{Makeup}, \text{Mustache}, \cdots\}$, and attributes irrelevant to $A$ as $X_r = \{\text{Age}, \cdots\}$ where $X$ is an image. To achieve the independence between $\mathbf{z}_d$ and $\mathbf{z}_r$, the total correlation is employed as in the previous works [6, 25] and approximated with adversarial training as $\mathcal{L}_{\text{TC}} = D_{\text{KL}}(q(a, \mathbf{z}_d, \mathbf{z}_r) || q(a, \mathbf{z}_d) q(\mathbf{z}_r))$. However, building graphs with specific $X_d$ and $X_r$ requires domain knowledge, and evaluating the validity of the graph design can be challenging and may not always be feasible. Moreover, not considering the causal relationship between $X_d$ and $X_r$ can impede learning independent latent codes.

### 3.4. Contradicting Assumption in Existing Methods

We claim that performance and fairness objectives in the previous works yield a contradiction, which can lead to potentially unstable optimization and unsatisfactory performance. We look into the disentanglement learning (Figure 1b) as an example, but this can be easily generalized to Figure 1a and 1c as discussed in Appendix.

In Figure 2, we summarized the data flow of disentanglement learning and the contradiction of two objectives. Fair disentanglement learning mainly has two goals: 1) predictiveness; 2) fairness. These methods project input $X$ into $\mathbf{z}_X$ and $\mathbf{z}_A$, where the predictiveness objective is to recover $Y$ and $A$ from $\mathbf{z}_X$ and $\mathbf{z}_A$, respectively. Commonly, cross-entropy $\mathcal{L}_{\text{CE}}$ is adopted, which is the lower bound of mutual information. If we achieve sufficient predictiveness by minimizing cross-entropy loss, *i.e.,* $\hat{Y} \approx Y$(*resp.* $\hat{A} \approx A$), it naturally results in high mutual information between the label $Y$ (*resp.* $A$) and latent code $\mathbf{z}_X$ (*resp.* $\mathbf{z}_A$). On the other hand, the fairness objective is to learn $\mathbf{z}_X$ and $\mathbf{z}_A$ that are independent by minimizing $\mathcal{L}_{\text{fair}}$, *i.e.,* $Z_X \perp Z_A$.

However, we argue that it is not possible to attain both objectives simultaneously due to one of the main causes of fairness problems in algorithmic decision-making, which is the undesired correlation between the target label $Y$ and sensitive information $A$ as highlighted in [39]. Under such data bias, the predictiveness objective yields $\hat{Y} \not\perp \hat{A}$. Accordingly, $\mathbf{z}_X$ and $\mathbf{z}_A$ cannot be independent, while the fairness objective requires independence. As a result, the two objectives contradict each other. We argue that overlooking the inherent data bias and forcing two subspaces to be independent may cause a poor fairness-accuracy trade-off [50] and unstable optimization.

## 4. FADES: Fair Disentanglement with Sensitive Relevance

Here, we propose a novel approach, FADES, to disentangle the features to learn fair representation. Unlike the previous works, we acknowledge that there exists an undesired correlation between $Y$ and $A$, and the related information cannot be explicitly partitioned. Instead, we propose to employ sensitive relevant code $\mathbf{z}_R$, which directs the information that overlaps between $Y$ and $A$. Here, $\mathbf{z}_R$ is responsible for

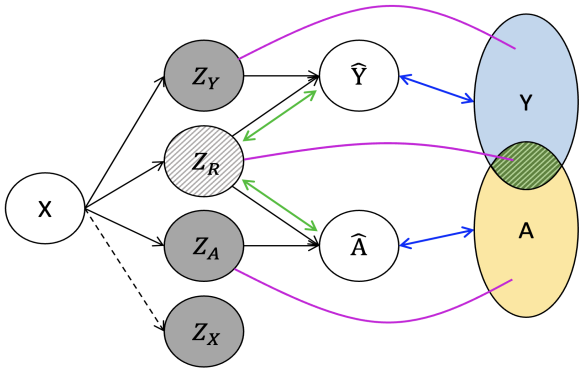



Figure 3. Illustration of the data flow of FADES. Since $\mathbf{z}_R$ is responsible for information correlated with $Y$ and $A$, we can achieve $\mathbf{z}_Y \perp \mathbf{z}_A$ along with perfectly recovering $Y$ and $A$. The purple lines connect the information in the observed space and their corresponding latent codes.

any correlated information between $Y$ and $A$. Thus, we can learn independent $\mathbf{z}_Y$ and $\mathbf{z}_A$ free of unwanted correlation while maintaining the predictiveness objective.

From the causal graph perspective as in Figure 1d, if $Y$ and $A$ are independent conditioned on $\mathbf{z}_R$, *i.e.,* $Y \perp A|\mathbf{z}_R$, then $\mathbf{z}_R$ is the only common cause of $Y$ and $A$. As a result, $\mathbf{z}_Y$ and $\mathbf{z}_A$ have zero mutual information, leading to independence. We illustrate the data flow and corresponding information projection in Figure 3. Then the fairness goal to achieve independence between $\mathbf{z}_Y$ and $\mathbf{z}_A$ by imposing independence between $\hat{Y}$ and $\hat{A}$ conditioned on $\mathbf{z}_R$, as the green arrows in Figure 3. $\mathbf{z}_X$ is an optional feature that is used to direct irrelevant features, such as the background of an image, for more controlled counterfactual generations.

### 4.1. Conditional Independence

We can first think of satisfying the proposed disentanglement by directly following the definition of conditional independence (CI):

$$p_\theta(\hat{Y}, \hat{A}|\mathbf{z}_R) = p_\theta(\hat{Y}|\mathbf{z}_R)p_\theta(\hat{A}|\mathbf{z}_R).$$

Since $\hat{Y}$ is predicted with $\mathbf{z}_R$ and $\mathbf{z}_Y$, we need to find the probability of prediction $Y$ conditioned only on $\mathbf{z}_R$, $p_\theta(Y = y|\mathbf{z}_R)$[1]. Empirically, we can compute the conditional probability of $k$-th sample by marginalizing over all $\mathbf{z}_Y$ within a batch similar to aggregate posterior as:

$$\begin{aligned} p_\theta(y|\mathbf{z}_R^{(k)}) &= \mathbb{E}_{p(\mathbf{x})}\left[\mathbb{E}_{q_\phi(\mathbf{z}_Y|\mathbf{x})}\left[p_\theta(y|\mathbf{z}_R^{(k)}, \mathbf{z}_Y)\right]\right] \\ &\approx \frac{1}{|B|}\sum_{i=1}^{B} p_\theta(y|\mathbf{z}_R^{(k)}, \mathbf{z}_Y^{(i)}), \end{aligned} \tag{2}$$

[1]With some abuse of notation, we omit the random variable in probability by denoting $p_\theta(Y = y|\mathbf{z}_R)$ as $p_\theta(y|\mathbf{z}_R)$ from here.

where $\mathbf{z}_R^{(i)}, \mathbf{z}_Y^{(i)}$ denotes the $i$-th sample in a mini-batch $B = \{\mathbf{z}_R^{(1)}, \cdots\}$ that are sampled uniformly from the distribution $\mathbf{x}^{(i)} \sim p(\mathbf{x}) = \frac{1}{|B|}$. To ensure independence, $p(\mathbf{z}_Y|\mathbf{z}_R) = p(\mathbf{z}_Y)$, we impose total correlation constraint as in Eqn (1) by factorizing $\{\mathbf{z}_Y, \mathbf{z}_R, \mathbf{z}_A\}$ with in-batch permutation. Similar to Eqn (2), we can compute

$$p_\theta(a|\mathbf{z}_R^{(k)}) \approx \frac{1}{|B|}\sum_{i=1}^{|B|} p_\theta(a|\mathbf{z}_R^{(k)}, \mathbf{z}_A^{(i)}).$$

Then we may minimize CI by minimizing the divergence $D(p_\theta(\hat{Y}, \hat{A}|\mathbf{z}_R)||p_\theta(\hat{Y}|\mathbf{z}_R)p_\theta(\hat{A}|\mathbf{z}_R))$. However, this computation is generally intractable.

### 4.2. Conditional Mutual Information

Instead, we propose to minimize *conditional mutual information* (CMI), $I_\phi(\hat{A}; \hat{Y}|\mathbf{z}_R)$, to learn such representation. In information theory, CMI for discrete random variables $X, Y$, and continuous random variable $Z$ is defined as:

$$\begin{aligned} I(X; Y|Z) &= \int_Z D_{\mathrm{KL}}(P_{(X,Y)|Z}||P_{X|Z} \otimes P_{Y|Z})dP_Z \\ &= \int_Z \sum_{y\in Y}\sum_{x\in X} P_{X,Y|Z}(x, y|z) \log \frac{P_{X,Y|Z}(x, y|z)}{P_{X|Z}(x|z)P_{Y|Z}(y|z)}. \end{aligned}$$

Proposition 4.1 below shows that minimizing CMI leads to conditional independence. The proof is in Appendix.

**Proposition 4.1.** *Conditional mutual information $I(X; Y|Z)$ is zero if and only if $X$ and $Y$ are independent conditioned on $Z$,* i.e., $X \perp Y|Z$.

Further, KL-divergence is lower bounded by total variation by Pinsker's inequality [7] as

$$\begin{aligned} &D_{\mathrm{KL}}(P_{(X,Y)|Z}||P_{X|Z} \otimes P_{Y|Z}) \\ &\qquad \geq \frac{1}{2}\sum_{X,Y}\left(\left|P_{(X,Y)|Z} - P_{X|Z} \otimes P_{Y|Z}\right|\right)^2. \end{aligned}$$

Therefore, reducing CMI, $I_\phi(\hat{A}; \hat{Y}|\mathbf{z}_R)$, is a valid objective to learn a representation that satisfies conditional independence $\hat{A} \perp \hat{Y}|\mathbf{z}_R$. Empirically, CMI loss is computed by utilizing the ground truth $Y$ and $A$ for training stability as:

$$\mathcal{L}_{\mathrm{CMI}} = \frac{1}{2}\left(I_\phi(\hat{A}; Y|Z_R) + I_\phi(\hat{Y}; A|Z_R)\right) \tag{3}$$

Then we can rewrite CMI with conditional entropy terms:

$$I_\phi(\hat{Y}; A|Z_R) = H_\phi(\hat{Y}|\mathbf{z}_R) - H_\phi(\hat{Y}|A, \mathbf{z}_R), \tag{4}$$

where $H_\phi(\cdot|\cdot)$ is conditional entropy. Then, we can empirically compute the first conditional entropy term in Eqn (4) by the following definition as

$$H_\phi(\hat{Y}|\mathbf{z}_R) = \frac{1}{|B|}\sum_{i=1}^{|B|}\sum_{y\in Y} -p_\theta(y|\mathbf{z}_R^{(i)}) \log p_\theta(y|\mathbf{z}_R^{(i)}), \tag{5}$$

where we can get $p_\theta(y|\mathbf{z}_R)$ from Eqn (2).

To compute the second conditional entropy, we need to compute $p_\theta(\hat{Y}|A, \mathbf{z}_R)$. This can be computed similarly for $\mathbf{z}_R^{(k)}$ sampled from an instance $\mathbf{x}^{(k)} \in B_a$ as

$$p_\theta(y|\mathbf{z}_R^{(k)}, a) = \mathbb{E}_{p(\mathbf{x}|A=a)}\left[\mathbb{E}_{q_\phi(\mathbf{z}_Y|\mathbf{x})}\left[p_\theta(y|\mathbf{z}_R^{(k)}, \mathbf{z}_Y^{(i)})\right]\right] \approx \frac{1}{|B_a|}\sum_{i=1}^{|B_a|} p_\theta(y|\mathbf{z}_R^{(k)}, \mathbf{z}_Y^{(i)}),$$

where $B_a$ denotes a subset of the batch with $A = a$. Then the conditional entropy can be computed as:

$$H_\phi(\hat{Y}|A, \mathbf{z}_R) = -\frac{1}{|B|}\sum_{a,y} p_\theta(y|a, \mathbf{z}_R)\log p_\theta(y|a, \mathbf{z}_R). \quad (6)$$

Now we can compute $\mathcal{L}_{\text{CMI}}$ in Eqn (3) by plugging Eqn (5) and (6). Note that $\mathcal{L}_{\text{CMI}}$ is not limited to binary $Y$ and $A$.

Note that when the whole information $Y \cup A$ is captured in $\mathbf{z}_R$, we can also achieve conditional independence or minimum CMI. To prevent this, we incorporate an information bottleneck regularization term as

$$\mathcal{L}_{\text{reg}} = -\big(H_\phi(\hat{Y}|\mathbf{z}_R) + H_\phi(\hat{A}|\mathbf{z}_R)\big), \quad (7)$$

which encourages $\mathbf{z}_R$ to encapsulate only the correlated information $Y \cap A$ by minimizing the confidence of prediction only given $\mathbf{z}_R$.

### 4.3. Final Objective Function of FADES

For our purpose, the ELBO objective can be rewritten as

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},y,a)}\big[\log p_\theta(\mathbf{x}|\mathbf{z}) + \log p_\theta(y|\mathbf{z}_Y, \mathbf{z}_R) + \log p_\theta(a|\mathbf{z}_A, \mathbf{z}_R)\big] - D(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})),$$

where $\mathbf{z} = [\mathbf{z}_X, \mathbf{z}_Y, \mathbf{z}_R, \mathbf{z}_A]$ represents the full latent space including the disentangled components. Then, the final objective function of FADES integrates $\mathcal{L}_{\text{ELBO}}$ with additional terms to enforce fairness and latent space independence as:

$$\mathcal{L}_{\text{FADES}} = -\mathcal{L}_{\text{ELBO}} + \lambda_{\text{CMI}}\mathcal{L}_{\text{CMI}} + \lambda_{\text{TC}}\mathcal{L}_{\text{TC}} + \beta\mathcal{L}_{\text{reg}}, \quad (8)$$

where $\lambda$ and $\beta$ are tunable hyperparameters.

## 5. Experiment

To provide a comprehensive comparison, we compare with at least one method from the class of graphical models presented in Figure 1, with the exception of correlation-aware learning, as it necessitates additional annotated data for constructing the causal graph, as discussed in Section 3.3. Accordingly, we mostly compare with recent fair disentanglement learning methods. Specifically, FairFactorVAE [32] and FairDisCo [31] are invariant learning methods which encode latent representation independent of sensitive attributes. FFVAE [6] is a disentanglement learning method minimizing the mutual information between the latent subspaces. GVAE [10] is a disentanglement learning method that aims at minimizing the leakage of unwanted information to learn fair representation. ODVAE [47] is a disentanglement learning method that learns fair representation without adversarial learning by enforcing the subspace to follow orthogonal priors. We set hyperparameters of all methods with a grid search to achieve the best EOD, especially for FADES, they are chosen in the range: $\lambda_{\text{CMI}}, \lambda_{\text{TC}} \in [0, 100]$ and $\beta \in [0, 1]$. All methods adopt ResNet-18 architecture [19] with 512 latent dimensions for vision tasks and 3 layered MLP encoder-decoder with 32 latent dimensions for the tabular dataset.

### 5.1. Fair Downstream Classification

The goal of fair classification is to minimize fairness violations while preserving predictive performance. To compare the methods, we conduct experiments on various fairness datasets. For facial attribute classification, we adopt CelebA [33] and UTK Face [56] datasets. Following previous works [49, 53, 55], the goal of CelebA is to predict "Smiling" attribute of a portrait, while for UTKFace, it is to classify whether a person in an image is over 35 years old, with gender as the sensitive attribute. Dogs and Cats dataset [8] is to distinguish a dog or cat given the color of its fur as the sensitive attribute. In addition, we evaluate on Adult income dataset [27], which is a popular fairness benchmark in structured data. The goal is to predict if the income of an instance exceeds \$50K with gender as the sensitive attribute.

For evaluation, we employ a 3-layered and 2-layered MLP classifier for vision and tabular datasets, respectively. The input for the classifier is the target-related features of pre-trained disentangled representation from each method. For instance, we fed $\mathbf{z}_Y$ of FADES and $\mathbf{z}_X$ of FFVAE [6], which is one of the disentanglement learning methods discussed in Section 3.2. For invariant learning methods, the whole latent space is utilized for downstream tasks. We measure the fairness violations with metrics including demographic parity (DP) [2] and equalized odds (EOD) [18]. For a fair comparison, we fixed the number of dimensions in the latent codes for all methods if applicable. We conduct 5 runs of experiments for each method with different random splits of the dataset if the split is not specified.

Table 1 summarizes the classification results on the benchmarks. Throughout the experiments, FADES demonstrated notable effectiveness in balancing classification accuracy with fairness measures. In particular, FADES consistently achieves the best accuracy while significantly improving the fairness violation. While FFVAE and ODVAE yield mixed results across the datasets, GVAE presents comparable results with those of FADES. Nonetheless, the proposed method consistently outperforms GVAE on all datasets at the same accuracy level. This validates the ef-

| | CelebA | | | UTKFace | | | Dogs and Cats | | | Adult | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc ↑ | EOD ↓ | DP ↓ | Acc ↑ | EOD ↓ | DP ↓ | Acc ↑ | EOD ↓ | DP ↓ | Acc ↑ | EOD ↓ | DP ↓ |
| FADES | 0.918 | **0.032** | 0.125 | 0.802 | **0.057** | **0.102** | **0.769** | 0.055 | 0.086 | 0.845 | 0.096 | 0.162 |
| GVAE [10] | **0.919** | 0.049 | 0.133 | **0.819** | 0.207 | 0.197 | 0.745 | 0.065 | 0.131 | **0.851** | 0.112 | 0.182 |
| FFVAE [6] | 0.891 | 0.075 | 0.071 | 0.767 | 0.271 | 0.206 | 0.727 | 0.067 | 0.111 | 0.802 | **0.063** | **0.092** |
| ODVAE [47] | 0.885 | 0.038 | 0.101 | 0.737 | 0.167 | 0.212 | 0.685 | **0.053** | **0.031** | 0.792 | 0.257 | 0.163 |
| FairDisCo [31] | 0.839 | 0.074 | **0.051** | 0.766 | 0.266 | 0.200 | 0.680 | 0.115 | 0.119 | 0.801 | 0.129 | 0.136 |
| FairFactorVAE [32] | 0.914 | 0.055 | 0.136 | 0.720 | 0.096 | 0.137 | 0.707 | 0.055 | 0.110 | 0.783 | 0.096 | 0.128 |

Table 1. Evaluation of downstream classification tasks on various benchmark datasets from learned representation.

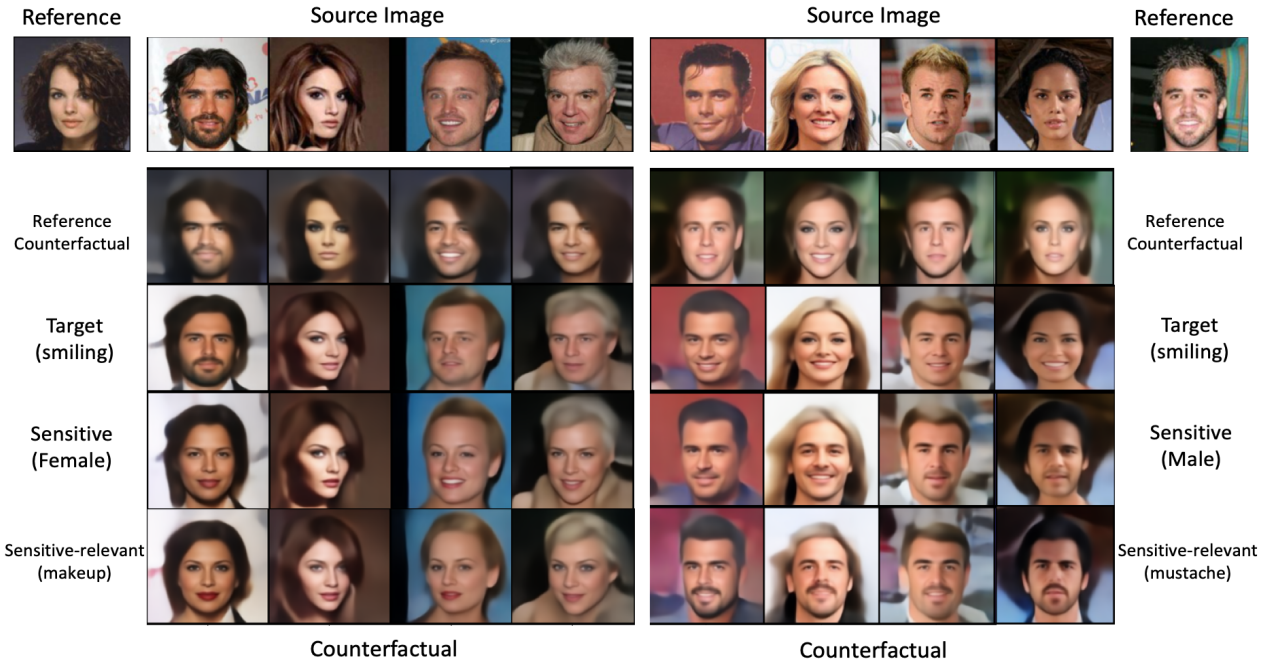



Figure 4. Illustration of counterfactuals (2-5 rows) given source and reference images on CelebA dataset in the first row. From the second row, we replace the latent subspaces $\mathbf{z}_X, \mathbf{z}_Y, \mathbf{z}_A$, and $[\mathbf{z}_A, \mathbf{z}_R]$ of the source images with those of the reference. Note that adding $\mathbf{z}_R$ for counterfactuals (Row 5) naturally adapts sensitive relevant features without domain knowledge, *e.g.,* mustache and makeup.

fectiveness of FADES in effectively disentangling correlated features, which leads to superior latent representations. Specifically, it implies the proposed method effectively isolated the features, preserving the quality information related to the target label $Y$ while simultaneously eliminating sensitive information $A$. Consequently, the learned representations ensure target-specific performance with no leakage of sensitive information with $\mathbf{z}_Y$.

In addition, we evaluate the needs of information bottleneck regularization $\mathcal{L}_{\text{reg}}$ in the objective of FADES as in Eqn (8). Figure 5 illustrates the results on 5 runs with varying $\beta \in [0, 1]$ and assessing the accuracy and EOD w.r.t. $\mathbf{z}_Y$. Notably, we observe proportional improvement in both accuracy and fairness violation as $\beta$ decreases until $\beta = 0.5$. This suggests that $\mathbf{z}_r$ may initially attempt to possess the comprehensive information to minimize CMI instead of distributing the information effectively. However, adequately setting $\beta$ mitigates the problem of information bottleneck and enables $\mathbf{z}_R$ to capture only correlated information. Conversely, the increase in fairness violation was noted at $\beta > 0.5$. We hypothesize that this phenomenon may be due to the overemphasis on maximizing the conditional entropy, inadvertently forcing the encoder to project

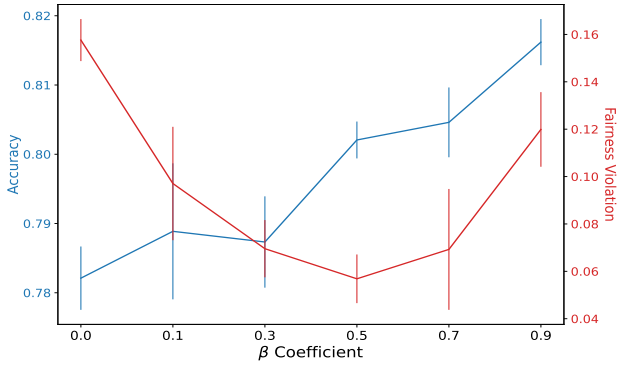



Figure 5. Trend of fairness and accuracy by sweeping information bottleneck weight $\beta$ on UTKFace dataset.

the correlated information to both $\mathbf{z}_Y$ and $\mathbf{z}_A$.

### 5.2. Fair Counterfactual Generation

We evaluate the counterfactual image generation task on CelebA dataset [33]. We consider *Smiling* as the target label and *Gender* as the sensitive attribute. Figure 4 depicts counterfactual generations by substituting portions of latent subspace with the ones from the reference images. The results showcase a smooth translation in the generations. Notably, $\mathbf{z}_R$ captures features semantically correlated with both $Y$ and $A$ without domain knowledge. Specifically, replacing $[\mathbf{z}_A, \mathbf{z}_R]$ shows a distinct difference with replacing only $\mathbf{z}_A$ by adding makeup (left half) and mustache (right half) to the image in the last row.

In Figure 6, we compare the generated image by linear interpolation of the sensitive code from a male with a mustache to another female sample. Each row depicts generated images when traversing sensitive code. The leftmost column is the reconstruction of a male with a mustache, and the rightmost column is when the sensitive code is substituted by a female sample. We observe that FADES provides better image quality than other methods when traversing the sensitive code. Specifically, FFVAE and FairDisCo generate a female with a mustache, which is less probable and unnatural. Additionally, GVAE and ODVAE had distortion on irrelevant features such as background color or hairstyle, which implies poor disentanglement. This is likely because it has no consideration of the sensitive relevant feature. On the other hand, FADES generates a much more natural female image while keeping other features unchanged. More result on feature translation is in Appendix.

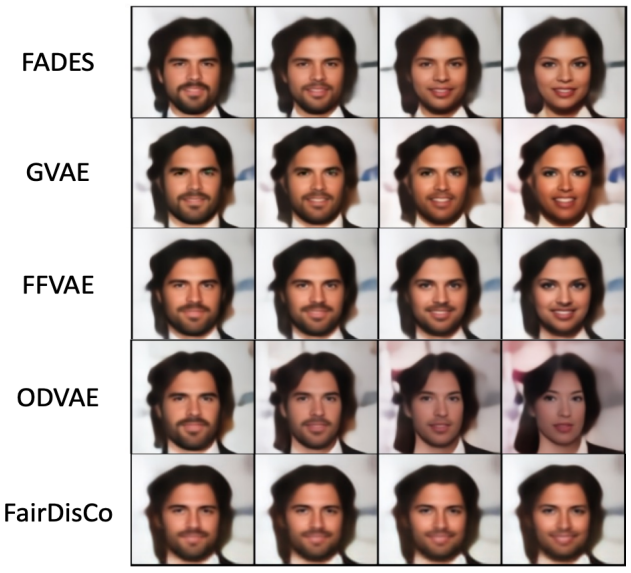



Figure 6. Reconstruction of interpolating sensitive code $\mathbf{z}_A$ starting from a male with a mustache to a female.

| | Acc (Digit)↑ | Acc (Color)↑ |
|---|---|---|
| FADES (Ours) | **93.46** | 95.31 |
| GVAE [10] | 85.98 | 88.20 |
| FFVAE [6] | 77.39 | 91.23 |
| ODVAE [47] | 77.45 | 83.26 |
| FairDisCo [31] | 87.21 | **99.31** |
| FairFactorVAE [32] | 87.33 | 90.18 |

Table 2. Digit and color recovery on unbiased C-MNIST.

### 5.3. Fair Image Reconstruction

We evaluate the fair image reconstruction on MNIST dataset with color bias [23, 41]. In the training set, 10 digits are correlated with 10 predefined colors with small random perturbation for 70% of samples, and 30% has uniform color assignment among the remaining colors. For the test set, every digit has uniform color assignments. The dataset is originally introduced to measure the color bias of a classifier predicting digits. In this experiment, we aim to investigate the ability of VAE-based models to disentangle the digit information color bias in the latent space. We train each method to reconstruct input images with color bias on the training set, and evaluate the reconstruction on the test set and expect to recover color and digit.

To quantitatively evaluate the reconstruction, we employ two pre-trained classifiers as oracles with accuracy over 99.7% on color and digit prediction tasks, respectively, trained on the unbiased dataset. We then measure the accuracy of recovering the digit and color by comparing the predicted digit and color of the reconstruction using the pre-trained classifiers with ground truth. In Table 2, we summarize the accuracy of digit and color recovery for comparing methods. We observe that most methods achieve higher accuracy on color than digit. In contrast, FADES renders both digit and color significantly better than others, indicating that the color and digit information has been effectively disentangled and recovered. Note that FairDisCo takes ground truth color label at decoder for reconstruction. See Appendix for more experiments.

## 6. Conclusion

In this work, we propose a novel disentanglement approach for addressing the limitations of existing fair methods under the prevailing data bias scenario. We claim that the naive assumption that the information is perfectly separable may lead to unstable training and performance drop. Instead, we direct the correlated information to a particular latent subspace by introducing sensitive relevant code. We theoretically demonstrate that we can effectively achieve fairness while maintaining performance without adversarial training by minimizing proposed conditional mutual information loss. Our experiments on fairness benchmarks validate the effectiveness of our proposed method. We discuss some limitations and future directions in the Appendix.

## References


[1] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *AAAI*, pages 2412–2420, 2019. 2

[2] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016. 6

[3] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018. 1, 2, 3

[4] Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7801–7808, 2019. 3

[5] Aidan Clark, Jeff Donahue, and Karen Simonyan. Adversarial video generation on complex datasets. *arXiv preprint arXiv:1907.06571*, 2019. 1

[6] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR, 2019. 1, 2, 3, 4, 6, 7, 8, 5

[7] Imre Csiszár and János Körner. *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press, 2011. 5

[8] Will Cukierski. Dogs vs. cats, 2013. 6

[9] Ayushman Dash, John Cristian Borges Gamboa, Sheraz Ahmed, Marcus Liwicki, and Muhammad Zeshan Afzal. Tac-gan-text conditioned auxiliary classifier generative adversarial network. *arXiv preprint arXiv:1703.06412*, 2017. 1

[10] Zheng Ding, Yifan Xu, Weijian Xu, Gaurav Parmar, Yang Yang, Max Welling, and Zhuowen Tu. Guided variational autoencoder for disentanglement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7920–7929, 2020. 1, 2, 3, 6, 7, 8, 4, 5

[11] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.*, 4(1):eaao5580, 2018. 1

[12] Mhairi Dunion et al. Conditional mutual information for disentangled representations in reinforcement learning. In *NeurIPS*, 2023. 3

[13] Cynthia Dwork. Differential privacy: A survey of results. In *International conference on theory and applications of models of computation*, pages 1–19. Springer, 2008. 1

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. 4

[15] Vincent Grari, Sylvain Lamprier, and Marcin Detyniecki. Adversarial learning for counterfactual fairness. *Machine Learning*, pages 1–23, 2022. 3

[16] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 3

[17] Laura Gustafson et al. Facet: Fairness in computer vision evaluation benchmark. In *ICCV*, 2023. 5, 6

[18] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323, 2016. 2, 6

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4

[21] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016. 1, 2, 3, 6

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1

[23] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019. 8

[24] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018. 2, 3, 6

[25] Hyemi Kim, Seungjae Shin, JoonHo Jang, Kyungwoo Song, Weonyoung Joo, Wanmo Kang, and Il-Chul Moon. Counterfactual fairness with disentangled causal effect variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8128–8136, 2021. 2, 3, 4

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1, 3

[27] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, pages 202–207, 1996. 1, 6

[28] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017. 2

[29] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020. 1

[30] Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR, 2021. 1

[31] Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. Fair representation learning: An alternative to mutual information. In *SIGKDD*, 2022. 2, 3, 6, 7, 8, 4, 5

[32] Shaofan Liu, Shiliang Sun, and Jing Zhao. Fair transfer learning with factor variational auto-encoder. *Neural Processing Letters*, 55(3):2049–2061, 2023. 2, 3, 6, 7, 8, 4, 5

[33] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 6, 8, 4

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 2

[35] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*, 30, 2017. 3

[36] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018. 2

[37] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 349–358, 2019. 3

[38] Emile Mathieu, Tom Rainforth, Nana Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pages 4402–4412. PMLR, 2019. 1

[39] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54 (6):1–35, 2021. 1, 2, 4

[40] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. *Advances in Neural Information Processing Systems*, 31, 2018. 3, 4

[41] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020. 1, 8, 5

[42] Changdae Oh, Heeji Won, Junhyuk So, Taero Kim, Yewon Kim, Hosik Choi, and Kyungwoo Song. Learning fair representation via distributional contrastive disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1295–1305, 2022. 4

[43] Or Patashnik et al. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 5

[44] Stephen R Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H Shah. Counterfactual reasoning for fair clinical risk prediction. In *Machine Learning for Healthcare Conference*, pages 325–358. PMLR, 2019. 3

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 5, 6

[46] Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2586–2594, 2019. 1

[47] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *European Conference on Computer Vision*, pages 746–761. Springer, 2020. 2, 3, 4, 6, 7, 8, 5

[48] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 5

[49] Zhibo Wang, Xiaowei Dong, Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, and Kui Ren. Fairness-aware adversarial perturbation towards bias mitigation for deployed deep models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2022. 6

[50] Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019. 4

[51] Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019. 3

[52] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575. IEEE, 2018. 2

[53] Tian Xu, Jennifer White, Sinan Kalkan, and Hatice Gunes. Investigating bias and fairness in facial expression recognition. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 506–523. Springer, 2020. 6

[54] Zheng Xu, Michael Wilber, Chen Fang, Aaron Hertzmann, and Hailin Jin. Learning from multi-domain artistic images for arbitrary style transfer. *arXiv preprint arXiv:1805.09987*, 2018. 1

[55] Huimin Zeng, Zhenrui Yue, Lanyu Shang, Yang Zhang, and Dong Wang. Boosting demographic fairness of face attribute classifiers via latent adversarial representations. In *2022 IEEE International Conference on Big Data (Big Data)*, pages 1588–1593. IEEE, 2022. 6

[56] Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5810–5818, 2017. 6

[57] Huaisheng Zhu and Suhang Wang. Learning fair models without sensitive attributes: A generative approach. *arXiv preprint arXiv:2203.16413*, 2022. 2, 3

[58] Wei Zhu, Haitian Zheng, Haofu Liao, Weijian Li, and Jiebo Luo. Learning bias-invariant representation by cross-sample mutual information minimization. In *ICCV*, 2021. 2