Enforcing Conditional Independence for Fair Representation Learning and Causal Image Generation

Jensen Hwa[1], Qingyu Zhao[2], Aditya Lahiri[3], Adnan Masood[4], Babak Salimi[3], Ehsan Adeli[1]
[1]Stanford University [2]Weill Cornell Medicine [3]University of California San Diego [4]UST
jphwa@cs.stanford.edu, eadeli@stanford.edu

Abstract

*Conditional independence (CI) constraints are critical for defining and evaluating fairness in machine learning, as well as for learning unconfounded or causal representations. Traditional methods for ensuring fairness either blindly learn invariant features with respect to a protected variable (*e.g.*, race when classifying sex from face images) or enforce CI relative to the protected attribute only on the model output (*e.g.*, the sex label). Neither of these methods are effective in enforcing CI in high-dimensional feature spaces. In this paper, we focus on a nascent approach characterizing the CI constraint in terms of two Jensen-Shannon divergence terms, and we extend it to high-dimensional feature spaces using a novel dynamic sampling strategy. In doing so, we introduce a new training paradigm that can be applied to any encoder architecture. We are able to enforce conditional independence of the diffusion autoencoder latent representation with respect to any protected attribute under the equalized odds constraint and show that this approach enables causal image generation with controllable latent spaces. Our experimental results demonstrate that our approach can achieve high accuracy on downstream tasks while upholding equality of odds.*

1. Introduction

Fairness in machine learning and computer vision is an increasingly important topic with growing applications to everyday life. The literature includes several methods for learning fair and unconfounded models based on domain adversarial invariant learning [4, 13, 18, 33, 37], statistical methods [16, 27], and information theory [21].

In the context of predictive modeling, algorithmic fairness aims to learn models that are unbiased towards protected subgroups by ensuring that the output label $\hat{y}$ is *invariant* to a sensitive attribute s. The existing plethora of fairness definitions typically captures this invariance in terms of a *conditional independence (CI)* (see Mehrabi et al. for a survey). For instance, the equalized odds criterion [10] requires prediction output $\hat{y}$ to be independent of the sensitive attribute s conditioned on the true class label y. We note that the conditional independence formulation $s \perp\!\!\!\perp \hat{y} \mid y$ is widely recognized as a more precise indicator of fairness compared with the marginal independence formulation $s \perp\!\!\!\perp \hat{y}$, due to the fact that some correlations between sensitive attributes and the target variable may be benign. CI plays a crucial role in capturing the true causal relationships

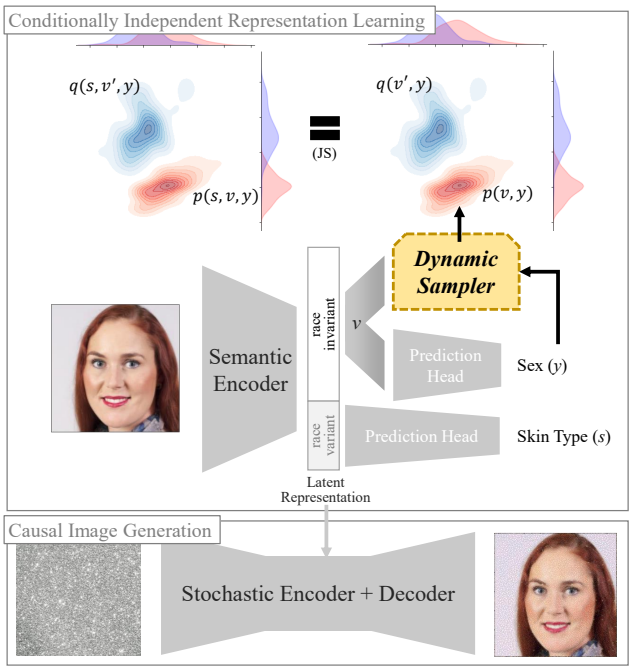

Figure 1. We propose a new way to ensure fairness in downstream tasks by enforcing conditional independence constraints on the latent representation. This is achieved by minimizing the Jensen-Shannon divergence (JS) between distributions obtained using a novel dynamic sampling technique. In the setting shown here, we apply our technique to the diffusion autoencoder's semantic representation to disentangle the sensitive attribute of skin type (a proxy variable for race) and perform causal image generation.

between the sensitive attributes and the target variable, ensuring that fairness is based on genuine causes of disparities rather than arbitrary associations [14, 24].

However, most existing work in fair machine learning typically focuses on enforcing independence on low-dimensional features in cases where the conditioning set is either empty or merely a low-dimensional categorical variable [2, 3, 24, 25, 37]. This restricts the learning of fair representations to limited settings involving only binary attributes in tabular datasets [35]. These methods cannot directly translate to high-dimensional spaces, such as the latent representations of large generative models.

Recently, Ahuja et al. developed a differentiable framework for enforcing CI in high-dimensional and continuous feature spaces by using a GAN-based approach in the context of data generation. To enforce $s \perp\!\!\!\perp \hat{y} \mid y$, they minimize the Jensen-Shannon (JS) divergence between the joint distribution $p(s, \hat{y}, y)$ and an auxiliary distribution $q(s, y', y)$ while the joint marginal $q(s, y)$ is distributed identically to $p(s, y)$ and $q(y' \mid y)$ is independent of s. Although this formulation enforces CI, it can only sample the joint probability distributions in the label space y and does not necessarily guarantee the learning of fair representations, *i.e.*, only the last network layer before predicting y removes the dependence while the rest of the network (including latent representations) remain biased or confounded.

Our approach introduces the concept of CI into the latent space by using a *dynamic sampler* to form the joint distributions with respect to the learned representations. This yields a computationally rigorous approach to fair and causal representation learning that exhibits a level of versatility not found in existing methods: We are able to control not just which features are encoded, but also where within the representation they appear. With this, we gain improved classification performance and, in the case of generative models such as the diffusion autoencoder (DiffAE), the ability to maintain fine-grained control over generated content. As a result, we can not only protect downstream tasks from bias against specific attributes, but also generate feature-invariant images corresponding to the enforced schema, allowing for a richer understanding of a model's representations.

2. Related Work

Invariant Representation Learning In the wake of increasingly biased large-scale models [1, 10, 18, 31], the learning of fair and unconfounded representations has taken on outsized importance. Methods based on domain adversarial learning are increasingly popular tools of choice in reducing bias and confounding effects [1, 15, 33, 34, 37]. Among these, Johndrow et al. and Tan et al. proposed methods for invariant feature learning and Zhao et al. presented models that minimize statistical mean dependence

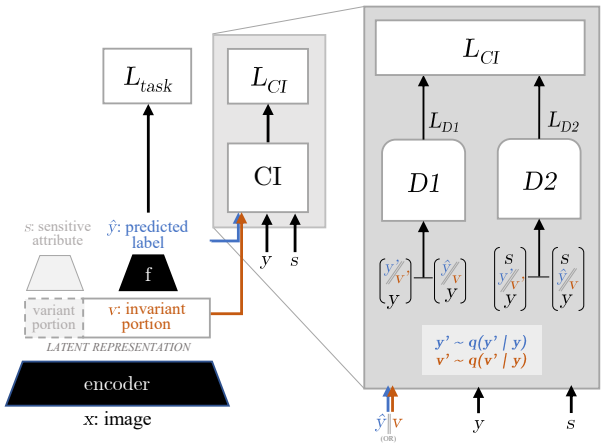

Figure 2. High-level view of our architecture. We introduce two variants of a conditional independence enforcer that can be added to any off-the-shelf encoder.

using a correlation-based adversarial loss function. Recently, Pirhadi et al. introduced a data cleaning method that enforces the conditional independence constraint for tabular data using optimal transport.

Other types of methods incorporating statistical operations have also been explored in the literature, such as using multivariate regression analysis [19] with general linear models [36]. Other approaches enforce fairness through post-processing steps on unfair trained classification models [6, 10, 36]. But since the training and fairness enforcement steps are conducted separately, these algorithms often lead to suboptimal fairness and accuracy trade-offs [32]. Recently, using traditional statistical methods, Lu et al. proposed a normalization layer that corrects the feature distributions with respect to labeled sensitive attributes. Their approach, known as metadata normalization, entailed a simple layer that could plug into any end-to-end model to protect the models against bias. Vento et al. then extended this work by turning the closed-form normalization operation into a network optimizable step. This is an active research field and many other approaches are being studied.

Conditional Independence for Algorithmic Fairness Consider a classifier or regression function $f(x)$ with output $\hat{y}$ and a *protected attribute* s, *e.g.*, sex or race. Algorithmic fairness aims to learn prediction functions or transformations that make the final outcome invariant or insensitive to protected attributes. Several widely used *associational* and *causal* notions of fairness, including the equalized odds criterion, can be attained by means of independence [25].

Rather than attempting to transform already trained unfair models to fair ones in a post-hoc manner, it is critical to enforce fairness constraints from the beginning of the model

training pipeline [11]. Apart from preventing the propagation of bias to downstream modeling tasks, this approach of fairness from a data management standpoint also leads to more robust and significant fairness measures by ensuring that data sources, transformations, and other training assumptions are sound [24].

Representation Learning with Diffusion Models Diffusion-based models learn to generate images by a denoising process: Random Gaussian noise is added to input images, and the model learns how to reverse this process to hallucinate new images from random noise. This family of models has proven to offer remarkable quality in image generation, promising to replace generative adversarial networks (GANs) [8] as the dominant architectural paradigm. Chief among them is the diffusion autoencoder [23], which encodes an image into a two-part latent subcode, capturing a stochastic representation via the aforementioned denoising approach and conditioning this process upon a semantic representation learned by a CNN. Interpolation of the resulting subcode results in smooth and meaningful changes in the decoded image. The semantic separation capability of DiffAEs has enabled various applications, such as attribute manipulation and various low-shot or zero-shot downstream applications [7, 9, 26]. In this work, we further enable DiffAEs to learn unbiased, causal, and conditionally independent latent subcode representations.

3. Method

Let $\{x, y, s\}$ be a training sample in the dataset, where x is an input image, y is the target prediction label, and s is a protected sensitive attribute. We aim to learn an encoder network $g_\theta(x) = v$ resulting in a latent representation v, from which a prediction network $f_\phi(v) = \hat{y}$ produces the final predicted label $\hat{y}$. With a slight abuse of notation, we use lower-case letters to also denote random variables. Ultimately, we aim to train a model that achieves high accuracy in predicting label y while ensuring equalized odds:

$$\begin{aligned} \underset{\theta,\phi}{\operatorname{argmin}} \quad & -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \\ \text{s.t.} \quad & s \perp\!\!\!\perp \hat{y} \mid y. \end{aligned} \tag{1}$$

We discuss two possible ways of enforcing the equalized odds fairness constraint: First, we enforce conditional independence with respect to the label y, *i.e.*, $s \perp\!\!\!\perp \hat{y} \mid y$. As discussed in Ahuja et al., this can be seen as a means to enforce equalized odds through CI. However, we claim that enforcing the independence constraint with respect to only the label is limited in its ability to enforce the equalized odds fairness constraint effectively. We therefore additionally propose to enforce the conditional independence with respect to the more information-rich latent representation v, or $s \perp\!\!\!\perp v \mid y$. We can then use this learned fair latent representation in a downstream prediction or generation task.

3.1. Enforcing CI with respect to Label

When the label y is highly correlated with the protected attribute s in the training data, a basic aim for a fair machine learning model is to ensure conditional independence of the predicted $\hat{y}$ from s, *i.e.*, $s \perp\!\!\!\perp \hat{y} \mid y$.

Assuming that the joint distribution $p(s, \hat{y}, y)$ is well-defined with respect to Lebesgue measure μ, previous work [3] demonstrates that the above conditional independence can be defined with respect to the Jensen-Shannon divergence (JS) between $p(s, \hat{y}, y)$ and an auxiliary distribution $q(s, y', y)$, where the joint marginal $q(s, y)$ is distributed identically to $p(s, y)$, and $q(y'|y)$ is independent of s. In other words, when $q(s, y', y) = p(s, y)q(y'|y)$, the conditional independence constraint $s \perp\!\!\!\perp \hat{y} \mid y$ is equivalent to enforcing

$$\text{JS}(p(\hat{y}, y), q(y', y)) = \text{JS}(p(s, \hat{y}, y), q(s, y', y)). \tag{2}$$

Accordingly, we can rewrite (1) in terms of JS as

$$\begin{aligned} \underset{\theta,\phi}{\operatorname{argmin}} \quad & -\frac{1}{n}\sum_{i=1}^{n} y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i) \\ \text{s.t.} \quad & \big(\text{JS}(p(\hat{y}, y), q(y', y)) - \\ & \text{JS}(p(s, \hat{y}, y), q(s, y', y)) \le \delta\big) \end{aligned} \tag{3}$$

for some $\delta > 0$ sufficiently small.

With the help of a conditional sampler $q(y' \mid y)$, the JS constraint of (3) can be satisfied by a general GAN architecture with two discriminators (Fig. 2) [3], *i.e.*, by minimizing

$$L_{\text{CI}} = (L_{\text{D1}} - L_{\text{D2}})^2 \tag{4}$$

where

$$\begin{aligned} L_{\text{D1}} &= \mathbb{E}[\log(1 - D_1(y', y)) + \mathbb{E}_{y,s}[\log D_1(\hat{y}, y)] \quad \text{and} \\ L_{\text{D2}} &= \mathbb{E}[\log(1 - D_2(y', s, y)] + \mathbb{E}_{y,s,y'}[\log D_2(\hat{y}, s, y)]. \end{aligned}$$

The key advantage of using (4) to derive conditional independence is that $q(y' \mid y)$ does not have to be a perfect sampler that exactly matches $p(\hat{y} \mid y)$. The only sufficient condition is that $q(y' \mid y)$ shares overlapping support with the original distribution $p(\hat{y} \mid y)$. This gives us many flexible ways to construct q, such as using a uniform distribution.

Specifically, in a supervised learning setting, where y is known at training time, we can set q using a uniform distribution over all raw model outputs that correspond to y. The encoder and prediction networks can then be learned by optimizing the loss function

$$L_{\text{Enc}} = \lambda L_{\text{CI}} + L_{\text{task}}, \tag{5}$$

where λ is a hyperparameter and L_{task} is the loss function for the prediction task. The discriminators can be trained adversarially to minimize the composite loss function:

$$L_{\text{D}} = L_{\text{D1}} + L_{\text{D2}}. \tag{6}$$

3.2. Enforcing CI with respect to Latent Representation

Enforcing conditional independence only with respect to $\hat{y}$ does not necessarily prevent the model from learning biased information from s in its representations. In this paper, we propose to enforce conditional independence with respect to the latent representation v in a way that most directly impacts the learnable parameters of the encoder. We achieve this by replacing the predicted label $\hat{y}$ with the latent vector v within our conditional independence loss described above, *i.e.*, $s \perp\!\!\!\perp v \mid y$. Similar to (2), our Jensen-Shannon divergence constraint becomes

$$\text{JS}(p(v, y), q(v', y)) = \text{JS}(p(s, v, y), q(s, v', y)). \tag{7}$$

As before, this requires a conditional sampler $q(v'|y)$. However, one problem that arises with this approach is that the conditional sampler is no longer well-defined at training time. Since the distribution and support of $p(v|y)$ is learned by the model on the fly, we cannot construct $q(v'|y)$ a priori.

To resolve this challenge, we implement the imperfect conditional sampler q via a novel **dynamic sampling** procedure. Specifically, to sample $v' \sim q(v'|y)$, we sample with replacement from the latent vectors associated with the given y within the same training batch. Similar to the bootstrapping procedure, our dynamic sampling can approximate the empirical distribution function $p(v|y)$, allowing the discriminators to keep up with and robustly train the encoder as the latent space is learned. Note that, other than the difference in the sampler, the objective function of (4) and the GAN architecture in Fig. 2 can be translated here by replacing $\{\hat{y}, y'\}$ with $\{v, v'\}$.

3.3. Disentangling DiffAE Latent Representation

The above construction of q provides us with a powerful technique that can be applied to any encoder architecture. We demonstrate this by using the state-of-the-art diffusion autoencoder model [23], which encodes images using a denoising diffusion process [26] conditioned upon a semantically meaningful representation inferred by a CNN encoder. An image, therefore, is encoded by two representations: a semantic subcode learned by the CNN, along with a stochastic subcode that, given the semantic subcode, denoises into the original image. The model learns to compress the most common high-level features into the semantic subcode and delegate the remaining, highly variable features to the stochastic representation. When applied to human face images, this yields a clean separation between anatomical features such as sex, facial structure, and skin type in the semantic subcode, and more transient qualities like pose, hairstyle, and expression in the stochastic subcode.

By enforcing conditional independence with respect to the semantic subcode of the diffusion autoencoder, we produce a latent representation that is invariant to a protected attribute of choice. The model adapts by encoding any facial features associated with the protected attribute in the stochastic subcode instead. When such facial features include the high-level attributes described earlier, this can unduly affect the natural dichotomy between the semantic and stochastic subcodes, restricting the expressive power of the semantic subcode and reducing the efficiency of the denoising process used for image generation. Empirically, this results in blurred and unrealistic generated images.

To overcome this, we enforce conditional independence with respect to only a portion of the semantic subcode. To further encourage the model to use the remaining portion to encode the facial features related to the sensitive attribute, we add a prediction head on the remaining portion to estimate the sensitive attribute s. We optimize this new prediction head alongside the encoder; instead of (5), we have

$$L_{\text{Enc}} = \lambda_1 L_{\text{CI}} + \lambda_2 \text{ BCE}(\hat{s}, s) + L_{\text{task}}. \tag{8}$$

In this way, the semantic subcode is apportioned into two parts: one variant to the sensitive attribute, and the other invariant. Although we only demonstrate this capability on tasks involving a single sensitive attribute, this approach can be easily extended to disentangle multiple attributes within the semantic subcode. With this formulation for causal disentanglement of the latent space, we show that the specific subcode in which a given facial attribute is represented can be methodically adjusted (*e.g.*, to change the race of the generated image), all while preserving the model's ability to accurately reconstruct other characteristics of the input image.

4. Experiments and Results

We apply our architecture to two different settings, each with a unique dataset, model, and confounding variable. In each case, we analyze the effectiveness of applying CI in the latent space as opposed to the label space and also compare with several baselines. Experiments were performed using a P100 GPU, with the exception of the diffusion autoencoder experiments which required four A100 GPUs.

Metrics We report prediction accuracies for each value of the protected attribute when s is discrete, along with the balanced accuracy (bAcc) to account for class imbalances. To evaluate invariance to the sensitive attribute, we use the squared distance correlation (dcor^2) [28] to quantitatively

measure the correlation between the protected variable and the learned features. When the protected variable is discrete, we also use the equality of opportunity (EO) independence metric, which measures the average gap in true positive rates for different possible values of the protected variable.

4.1. Synthetic Experiments

As an initial proof of concept, we train a convolutional neural network on synthetically generated image data. Each image in this dataset is of size 32×32 and composed of four Gaussian kernels, as shown in Fig. 3. The diagonal kernels are of equal intensity σ_A, and the off-diagonal kernels are of equal intensity σ_B. For half of the images, we sample $\sigma_A, \sigma_B \sim \mathcal{U}(1,4)$ and assign the label $Y = 0$. For the other half, we sample $\sigma_A, \sigma_B \sim \mathcal{U}(3,7)$ and assign $Y = 1$. We then denote σ_B as the protected attribute and analyze the model's ability to predict an image's label based on the diagonal kernels controlled by σ_A while ignoring the off-diagonal kernels controlled by σ_B. We do this by enforcing the conditional independence condition separately with respect to the label space ($\sigma_B \perp\!\!\!\perp \hat{Y} \mid Y$) and the latent space ($\sigma_B \perp\!\!\!\perp V \mid Y$).

Whereas an unconstrained model may achieve maximum theoretical accuracy of $1 - \frac{1}{2}\left(\frac{1}{3}\right)^2 = 0.94$, correctly predicting all cases except half of those where $\sigma_A, \sigma_B \in (3,4)$, a fair classifier using only the information represented by σ_A can attain a theoretical accuracy of $1 - \frac{1}{2}\left(\frac{1}{3}\right) = 0.83$.

Our encoder is a simple CNN comprised of a convolutional layer, ReLU activation, max pooling, convolution, ReLU, and two linear layers, applied in that order. The first linear layer has an output of length 10, which we define as the latent space v, while the second linear layer converts this latent space into the single output logit corresponding to y.

In Table 1, we report metrics after training each model using a batch size of 512 and a learning rate of $1\text{e}{-4}$. We select the λ hyperparameter by 5-fold cross-validation and report metrics based on a separate test set.

Fig. 5 demonstrates the tradeoff of fairness versus accuracy and dcor^2 for both model variants. The latent space

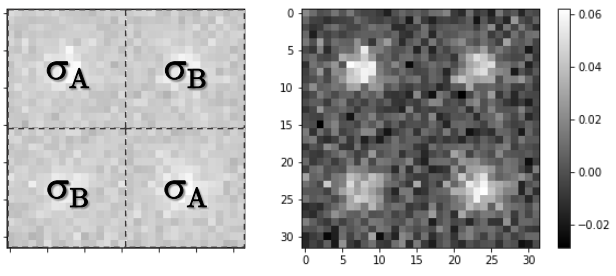

Figure 3. Synthetic data format and sample. The diagonal kernels are controlled by σ_A, while the off-diagonals are controlled by σ_B.

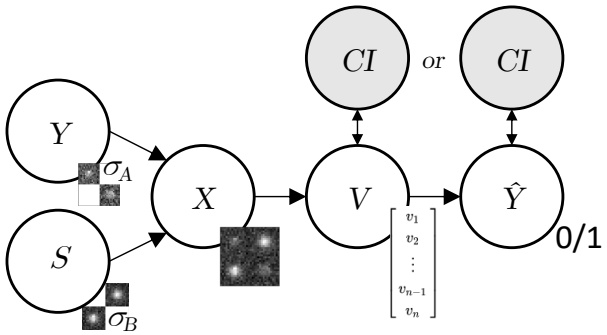

Figure 4. Synthetic data causal diagram. We apply the L_{CI} component to either the latent vector space (V) or the label space (Y).

Table 1. Synthetic data experiment results. Balanced accuracy (bAcc) closer to 0.83 is better, as is lower dcor^2.

Model	**bAcc**	**dcor^2**
Vanilla CNN	0.94	0.432
Regularized v-space CNN	0.80	0.382
y-space CI-CNN	0.87	0.298
v-space CI-CNN	**0.84**	**0.055**

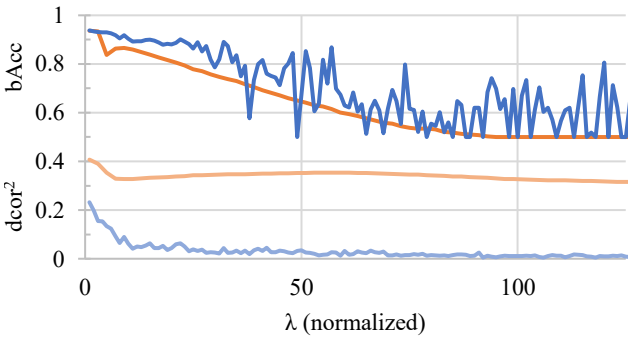

Figure 5. Fairness and accuracy metrics versus conditional independence strength λ. Orange lines correspond to the y-space CI-CNN, and blue lines to the v-space CI-CNN.

variant introduces some instability for larger λ, but achieves increased accuracy alongside decreased dcor^2. While some instability is present for large λ, this largely does not appear until after dcor^2 has converged.

In Fig. 6, we show each model's unnormalized output for each integer combination of σ_A and σ_B. Enforcing CI in the latent space produces outputs that remain nearly unaffected by changes in the protected attribute σ_B, as evidenced by near-constant values within each row.

Model B1: Vanilla CNN As a baseline, we train the standalone CNN without any conditional independence enforcing mechanism, using only a binary cross entropy loss function. This model relies heavily on the protected attribute

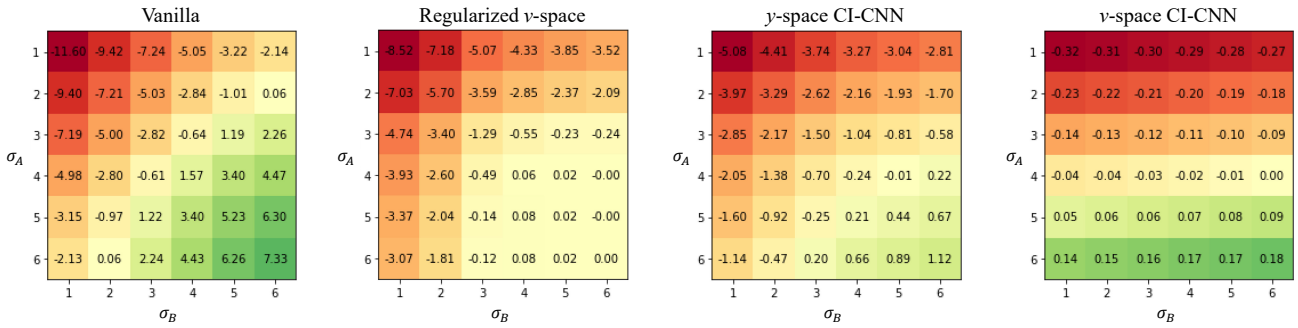

Figure 6. Unnormalized model predictions on synthetic data for given σ_A, σ_B. Negative values imply a prediction of $\hat{y} = 0$, while positive values correspond to $\hat{y} = 1$.

σ_B, as shown by a high bAcc and $dcor^2$.

Model B2: Regularized v-space CNN To evaluate the efficacy of our conditional independence enforcement, we establish a baseline by embedding the equalized odds condition directly into the loss function. Defining R_0 and R_1 to be the absolute difference between the true positive rate and false positive rate for $Y = 0$ and $Y = 1$ respectively, the loss function becomes

$$L = \lambda(R_0 + R_1) + \text{BCE}(\hat{y}, y). \tag{9}$$

We tune the λ parameter over the validation set to target the theoretical maximum accuracy under conditional independence. This regularizer appears to have an overbearing effect on the model, and despite the reduced dcor^2, it struggles to learn a meaningful representation.

Model 1: y-space CI-CNN We use the output of the final linear layer as input to our conditional independence enforcing loss component and apply binary cross entropy loss:

$$L_{\text{Enc}} = \lambda L_{\text{CI}} + \text{BCE}(\hat{y}, y), \tag{10}$$

where, as in (5), λ controls the strength of conditional independence enforcement and L_{CI} is the loss from the dual discriminators. We find that the model learns a representation somewhat uncorrelated to σ_B, but only in an unstable manner without much improvement upon the regularized baseline.

Model 2: v-space CI-CNN In our second iteration, we insert the conditional independence component closer to the latent space of the model, before the final linear layer, and employ the dynamic sampling procedure to simulate the conditional latent space distribution. The resulting model exhibits a much smaller confounding effect, achieving an accuracy closest to the theoretical maximum under CI and the lowest correlation between the protected and target variables. This demonstrates the effectiveness of our dynamic resampling procedure, which makes practical the enforcement of conditional independence with respect to the latent space and is one of the key contributions of this paper.

4.2. Face Image Experiments

In selecting an experimental setting in which to apply our architecture to the diffusion autoencoder, we searched for datasets (1) labeled with a relevant secondary attribute that we could treat as a confounding variable, and (2) compatible with other public datasets that could be used to pretrain the DiffAE before finetuning. One of the few datasets meeting these criteria was a set of 1,270 face images provided by the Gender Shades project [5], representing subjects from 3 African countries and 3 European countries. Each image is labeled with both skin type, using the Fitzpatrick classification system, and sex, inheriting a binary male/female grouping as a simplified proxy for gender. We aligned and cropped the images to the format expected by the diffusion autoencoder model, and then employed our architecture to train the diffusion autoencoder to predict sex while being conditionally independent to skin type.

We initialized a pretrained diffusion model based on the 128×128 Flickr-Faces-HQ Dataset, which consists of 70,000 highly varied face images. We treat the model's semantic subcode as the latent space and add a single linear layer prediction head to convert from this 512-length vector space to a single logit representing the sex label space.

Models B1-3 We implement three baseline models. BR-Net [1] uses adversarial training to learn image features that have statistical "mean independence" with the protected attribute. The adversarial objective of BR-Net is based on Pearson's correlation between the true and the predicted value of the protected attribute. Second, Kim et al. uses mutual information minimization to ensure that the learned image features cannot predict the protected attribute. Fi-

Table 2. Classification of sex from Gender Shades dataset facial images, with mean and standard deviation across five runs.

Model	**bAcc (%)**	I	II	III	IV	V	VI	**EO (%)**	**dcor**^{2}
Kim *et al.* [13]	88.8 ± 1.3	87.8	93.0	93.6	91.7	87.4	86.2	8.4 ± 3.6	**0.022** ± 0.012
Multi-task [17]	93.9 ± 0.4	92.7	95.6	93.6	100	95.8	87.3	7.3 ± 1.6	0.190 ± 0.012
BR-Net [1]	94.8 ± 0.4	92.7	97.4	93.6	100	95.8	88.5	***5.3*** ± 1.2	0.170 ± 0.007
Vanilla Diffusion Autoencoder	93.0 ± 0.6	85.4	95.6	100	91.7	87.3	90.1	11.5 ± 1.1	0.326 ± 0.000
y-space CI-DiffAE	***94.8*** ± 2.1	80.5	93.7	93.6	87.5	81.1	86.2	9.9 ± 6.0	0.120 ± 0.034
v-space CI-DiffAE	**96.6** ± 1.8	97.5	100	97.9	91.7	93.7	87.3	**5.0** ± 2.5	***0.076*** ± 0.025

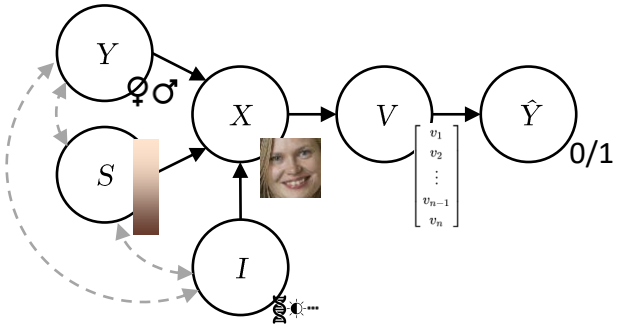

Figure 7. Causal graph corresponding to face dataset. By way of the latent space V, we train a model to predict sex (Y) from face image (X) while being invariant to skin color (S). Here, countless other factors (I) such as genetics and lighting also influence each image, and hidden confounders (dashed lines) cause correlation among inputs.

nally, a multi-task network [17] is implemented to use the same encoder to predict both the target value and the protected attribute simultaneously.

Model B4: Vanilla Diffusion Autoencoder For a closer comparison to our architecture, we train the diffusion autoencoder using its default loss function, denoted in [23] as L_{simple}. In a separate optimization step, we train the prediction head using binary cross entropy loss (BCE).

Model 1: y-space CI-DiffAE We unify the training of the diffusion autoencoder and the prediction head under a single loss function and enforce conditional independence with respect to the output of the prediction head. In the context of (5), the encoder objective function becomes

$$L_{\text{Enc}} = \lambda L_{\text{CI}} + (L_{\text{DiffAE}} + \rho\, \text{BCE}(y, \hat{y}))\,, \qquad (11)$$

for hyperparameter ρ.

Model 2: v-space CI-DiffAE In an analogous manner to the v-space CI-CNN, we enforce conditional independence directly on the semantic subcode, keeping other portions of the model the same.

Results in Table 2 indicate that the v-space constraint yields an all-around improvement in accuracy and fairness compared to the vanilla model, as opposed to the y-space constraint which achieves lower dcor^2 with considerable impact on accuracy. This improvement is statistically significant as measured by a McNemar test ($\chi^2 = 8.257$, p-value $= 0.0040$). Overall, these results demonstrate that the v-space CI constraint produces results well on-par with existing work. Moreover, this technique has the added capability of causal image generation, described below.

4.2.1 Visual Results

Due to the strong association between skin type and race, skin type impacts a large number of facial features. Our architecture removes much of this association, which creates a race-invariant representation, but as a result, also constricts the model's ability to encode anatomical facial features in the semantic subcode. Therefore, nearly all information becomes encoded by the stochastic subcode, and generated images are less realistic.

As discussed previously, we remedy this effect by enforcing conditional independence on only a portion of the semantic subcode and simultaneously training a sensitive attribute prediction head on the remaining portion. We find that this partial CI-DiffAE architecture allows the semantic subcode as a whole to continue encoding the facial features necessary to produce a clean image. To interpret the effects on the semantic subcode, we generate images after adjusting the race-variant portion of the semantic subcode to values corresponding to each skin type (Fig. 8). We observe that the v-space model is effectively (and causally) able to disentangle race-related features within the semantic subcode, resulting in a smooth and meaningful transition with changes limited to those relating to race.

Furthermore, by sampling the race-invariant portion of the semantic subcode, we retain the diffusion autoencoder's ability to generate novel images, but with added control over skin type (Fig. 9). These images demonstrate how our technique creates the race-invariant portion of the latent representation by removing the direct causal effect of skin shade, analogous to deleting the solid arrow between S and

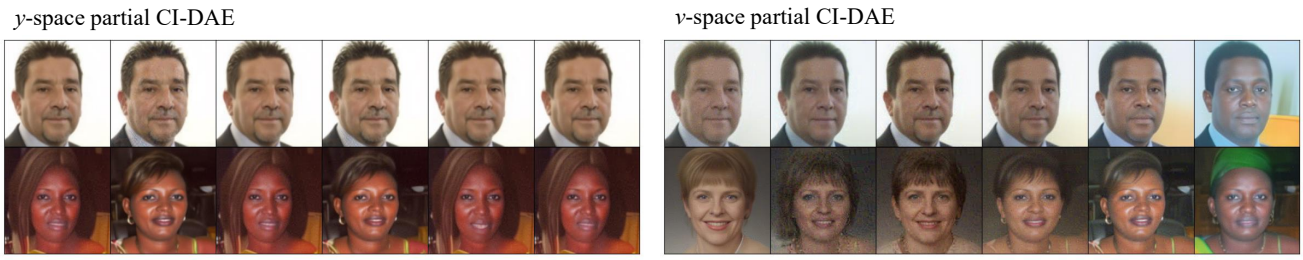

Figure 8. Selected image reconstructions. For each variant of our CI-DiffAE, we reconstruct two images from the dataset while adjusting the race-variant portion of the semantic subcode: the leftmost image in each group corresponds to skin type 1, and the rightmost to skin type 6. The latent (v-space) CI constraint can effectively disentangle skin shade from other facial features.

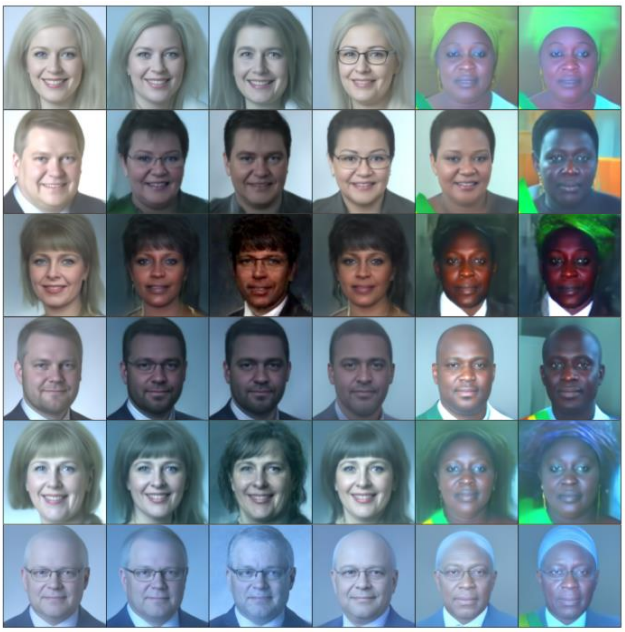

Figure 9. Race-invariant image generation using our v-space partial CI-DiffAE architecture. Each column contains hallucinated images of the same skin type.

X in Fig. 7. At the same time, due to the fine-grained nature of the equalized odds criterion we enforce, the representation preserves the indirect effect of skin shade that is mediated by sex, shown in Fig. 7 as the dashed arrow between Y and S. Therefore, the model continues to capture differences in gender expression across race, most notably in the headwraps often worn by women in Sub-Saharan African cultures. By disentangling race from the latent representation, we gain a richer insight into the causal processes that underlie facial images, crucial for effective image generation.

5. Conclusion

We introduced a framework to ensure fair and unconfounded representation learning during training and demonstrated both its versatility when applied to complex models and its effectiveness when compared to alternative methods. We iterated upon the theoretical idea of expressing the conditional independence constraint as an equality of two Jensen-Shannon divergences and extended this to high dimensional latent space via a dynamic sampling technique that can be easily implemented for any encoder. Our work exposes a new approach to generally enforce a conditional independence constraint on a model, which can then be used in downstream tasks such as causal image generation and fair predictive models. To our knowledge, this is the only model-agnostic training approach to be shown effective on enforcing specific features to be encoded in given dimensions of the latent space. We are optimistic that further experimentation will reveal applications to other tasks concerning fairness and disentanglement. Moreover, as a direction for future work, the use of conditionally invariant embeddings may prove useful in extending traditional causal methods like mediational analysis to complex, high-dimensional settings.

Acknowledgement

This research was partially supported by UST, Stanford Institute for Human-Centered AI (HAI) GCP cloud credits, and National Institutes of Health (NIH) grants U54HG012510 and AG084471. J.H. was also supported by a research fund from Panasonic and E.A. by the Stanford School of Medicine Jaswa Innovator Award.

References

[1] Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V. Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M. Pohl. Representation learning with statistical independence to mitigate bias. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2512–2522, 2021. 2, 6, 7

[2] Raj Agrawal, Chandler Squires, Neha Prasad, and Caroline Uhler. The decamfounder: Non-linear causal discovery in the presence of hidden variables, 2021. 2

[3] Kartik Ahuja, Prasanna Sattigeri, Karthikeyan Shanmugam, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Murat Kocaoglu. Conditionally independent data generation. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 2050–2060. PMLR, 2021. 2, 3

[4] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Adversarial invariant feature learning with accuracy constraint for domain generalization. *arXiv preprint arXiv:1904.12543*, 2019. 1

[5] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018. 6

[6] Michael Feldman. *Computational fairness: Preventing machine-learned discrimination*. PhD thesis, Haverford College, 2015. 2

[7] Giorgio Giannone, Didrik Nielsen, and Ole Winther. Few-shot diffusion models. *arXiv preprint arXiv:2205.15463*, 2022. 3

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[9] Jiechao Guan, Manli Zhang, and Zhiwu Lu. Large-scale cross-domain few-shot learning. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 3

[10] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323, 2016. 1, 2

[11] Maliha Tashfia Islam, Anna Fariha, Alexandra Meliou, and Babak Salimi. Through the data management lens: Experimental analysis and evaluation of fair classification. In *Proceedings of the 2022 International Conference on Management of Data*, pages 232–246, 2022. 3

[12] James E Johndrow, Kristian Lum, et al. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 2019. 2

[13] 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 Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019. 1, 6, 7

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

[15] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 2

[16] Mandy Lu, Qingyu Zhao, Jiequan Zhang, Kilian M Pohl, Li Fei-Fei, Juan Carlos Niebles, and Ehsan Adeli. Metadata normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10917–10927, 2021. 1, 2

[17] Yongxi Lu, Abhishek Kumar, Shuangfei Zhai, Yu Cheng, Tara Javidi, and Rogerio Feris. Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1131–1140, 2017. 7

[18] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*, 2018. 1, 2

[19] Roseanne McNamee. Regression modelling and other methods to control confounding. *Occupational and environmental medicine*, 62(7):500–506, 2005. 2

[20] 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

[21] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems 31*, pages 9084–9093. Curran Associates, Inc., 2018. 1

[22] Alireza Pirhadi, Mohammad Hossein Moslemi, Alexander Cloninger, Mostafa Milani, and Babak Salimi. Otclean: Data cleaning for conditional independence violations using optimal transport. *SIGMOD*, 2024. 2

[23] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10619–10629, 2022. 3, 4, 7

[24] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the 2019 International Conference on Management of Data*, pages 793–810, 2019. 2, 3

[25] Babak Salimi, Bill Howe, and Dan Suciu. Database repair meets algorithmic fairness. *ACM SIGMOD Record*, 2020. 2

[26] Abhishek Sinha, Jiaming Song, Chenlin Meng, and Stefano Ermon. D2c: Diffusion-decoding models for few-shot conditional generation. *Advances in Neural Information Processing Systems*, 34:12533–12548, 2021. 3, 4

[27] Chandler Squires, Dennis Shen, Anish Agarwal, Devavrat Shah, and Caroline Uhler. Causal imputation via synthetic interventions, 2020. 1

[28] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007. 4

[29] Zilong Tan, Samuel Yeom, Matt Fredrikson, and Ameet Talwalkar. Learning fair representations for kernel models. In *International Conference on Artificial Intelligence and Statistics*, pages 155–166. PMLR, 2020. 2

[30] Anthony Vento, Qingyu Zhao, Robert Paul, Kilian M Pohl, and Ehsan Adeli. A penalty approach for normalizing feature distributions to build confounder-free models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 387–397. Springer, 2022. 2

[31] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019. 2

[32] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017. 2

[33] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 1, 2

[34] Zhongwei Zhang, Mingyu Shao, Chicheng Ma, Zhe Lv, and Jilei Zhou. An enhanced domain-adversarial neural networks for intelligent cross-domain fault diagnosis of rotating machinery. *Nonlinear Dynamics*, 108(3):2385–2404, 2022. 2

[35] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J Gordon. Conditional learning of fair representations. *arXiv preprint arXiv:1910.07162*, 2019. 2

[36] Qingyu Zhao, Ehsan Adeli, Adolf Pfefferbaum, Edith V Sullivan, and Kilian M Pohl. Confounder-aware visualization of convnets. *arXiv preprint arXiv:1907.12727*, 2019. 2

[37] Qingyu Zhao, Ehsan Adeli, and Kilian M Pohl. Training confounder-free deep learning models for medical applications. *Nature communications*, 11(1):1–9, 2020. 1, 2