INTRODUCTION

The key assumption behind standard machine learning techniques is that training and test data are independently and identically distributed (i.i.d.). In practice, different forms of bias can creep into our datasets in subtle ways, breaking this assumption [Torralba and Efros 2011], [Tommasi et al. 2017], [Ritter et al. 2017], [Stock and Cisse 2018], [Grover et al. 2020]. As a result, machine learning systems can become worthless or even harmful. For instance, a classifier may learn to make decisions based on a confounder present in the training data and then fail to generalize to the test data [Lapuschkin et al. 2019]. Similarly, a machine learning system may propagate unfair biases present in the training data by learning to make decisions based on protected attributes such as race or gender [Hardt et al. 2016].

A number of methods have been proposed to repair deep classifiers which have learned the wrong rules, or to prevent them from doing so in the first place [Ross et al. 2017], [Murdoch et al. 2018], [Ieso and Kersting 2019], [Rieger et al. 2019], [Efron et al. 2019], [Selvaraju et al. 2019], [Schramowski et al. 2020], [Shao et al. 2021], [Kim et al. 2019], [Grover et al. 2019]. Generative models however have generally not been addressed, despite the fact that they have the potential to amplify bias by generating more biased data at test time [Grover et al. 2020]. In this work, we confirm experimentally that generative models are prone to learning biases as well. A commonly used approach to combat this issue is data augmentation. When it is easy to generate synthetic data in which the confounding variable is randomly varied, then an augmented version of the original training set can be constructed which no longer exhibits (undesirable) bias [Geirhos et al. 2019]. However, when such data needs to be manually collected, this approach is typically prohibitively expensive, as the number of ways in which factors of variation can be combined grows exponentially [Bahng et al. 2020].

In this paper, we show how this issue can be prevented by learning disentangled representations. A representation is called disentangled if its variables correspond directly to the factors of variation underlying the data [Bengio et al. 2013], [van Steenkiste et al. 2019], [Locatello et al. 2019a], [Creager et al. 2019], [Chartsias et al. 2019], [Higgins et al. 2017]. Specifically, we study variational autoencoders (VAEs) [Kingma and Welling 2014]. When such a model is presented with biased data, if no precautions are taken, it will generally match the biased training distribution and learn a representation in which the underlying factors of variation are entangled. If we can however disentangle these factors, we can place an independent prior over the latent variables and thereby force the models to learn a distribution in which the factors of variation are independent. This is a significant challenge, since in addition to achieving disentanglement,
Figure 1: **Left:** Our model is trained on a biased version of the colored MNIST dataset, in which color and shape are fully correlated (diagonal marked in orange). As human feedback, we supply a small number of labelled examples from the first row and the rightmost column. **Right:** Despite never observing most combinations of color and shape (observed combinations are marked in orange and green), our model learns to generate all of them. Here, reconstructions are drawn by combining the shape (top row) with the color of the digit (leftmost column).

such a model needs to generalize to those combinations of attributes which were not present in the training data due to bias.

To learn such disentangled representation, we use a small amount of human feedback. The feedback technique is built on recent work on weakly supervised disentanglement [Chen and Batmanghelich, 2020, Locatello et al., 2020, Shu et al., 2020]. Specifically, it takes the form of a small amount of labelled data in which one attribute remains constant, while the others are varied. Even though most combinations are never observed by the model, it nevertheless learns to generate, reconstruct, and encode such examples correctly, as illustrated in Fig. 1. We find that generalization can also be achieved with this small amount of feedback. This addresses the often prohibitive cost of data augmentation.

Our key contributions are: (1) We propose a method for learning unbiased, disentangled representations from severely biased data by using human feedback within the VAE framework. (2) We formulate human feedback as a set of examples with partially observed labels to disentangle the internal representations of VAEs. (3) We prove that this approach can guarantee restrictiveness and non-trivial solutions when disentangling a subset of factors, accommodating for possible nuisance factors. (4) We demonstrate empirically on several benchmarks that our unbiased generative model generalizes to novel combinations of the underlying factors of variation. The disentangled latent representation allows fine-grained control over the generated samples.

We proceed by touching upon related work. Then, we introduce our approach of learning disentangled representations via human feedback and provide some theoretical guarantees. Before concluding, we present the results of our empirical evaluation.

### 2 Related Work

**Correcting Machine Learning Models.** Machine learning models trained on biased data may exhibit undesirable bias due to data-collecting process. A number of works have been done to correct machine learning models from learning this bias [Ross et al., 2017, Murdoch et al., 2018, Teso and Kersting, 2019, Rieger et al., 2019, Erion et al., 2019, Selvaraju et al., 2019, Schramowski et al., 2020, Shao et al., 2021]. However, these works can only remove very simple bias on the observational feature level. What if the bias is on a more abstract level? [Kim et al., 2019] propose to unlearn more complex bias in a data set by minimizing the mutual information between the transformed feature and the target bias. These works are all restricted to discriminative models.

**Disentangled Representation.** In generative modelling, learning disentangled representations is an active area of research that receives increasingly attention [Higgins et al., 2016, Kumar et al., 2017, Kim and Mnih, 2018, Chen et al., 2018, Burgess et al., 2018, Kim and Mnih, 2018, Mathieu et al., 2019]. $\beta$-VAE [Higgins et al., 2016] is one of the earliest work of this kind, which augmented the lower bound formulation to regulates the strength of independence prior pressures. FactorVAE [Kim and Mnih, 2018] encourages the distribution of representations to be factorial and hence independent across the dimensions. $\beta$-TCVAE [Chen et al., 2018] proposes an equivalent objective as Factor-
VAE but optimize it differently. However, these unsupervised learning of disentangled representations is shown to be fundamentally impossible from i.i.d. observations without inductive biases or some form of supervision [Locatello et al. 2019b]. Thereafter this research problem starts to be addressed with supervision to provide general solutions. Locatello et al. [2019b], Shu et al. [2020], Sorrenson et al. [2020], Khemakhem et al. [2020], Chen and Batmanghelich [2020] use weak supervision by providing similarities between instances based on a factor to be disentangled and formulate consistency and restrictiveness. Locatello et al. [2020] also use paired observations as weak supervision and they require disentanglement guarantees of weak supervision algorithms. In particular, they decompose disentanglement into consistency and restrictiveness. Shu et al. [2020] propose a theoretical framework to analyze the use weak supervision by providing similarities between instances based on a factor to be disentangled and formulate a regularized ELBO to enforce disentanglement. Shu et al. [2020] propose a theoretical framework to analyze the disentanglement guarantees of weak supervision algorithms. In particular, they decompose disentanglement into consistency and restrictiveness. Locatello et al. [2020] also use paired observations as weak supervision and they require weaker assumptions than Shu et al. [2020], Locatello et al. [2019c] use a small number of labels to learn disentangled representations.

3 Unbiased Generative Models via Disentanglement

We formalize the debiasing setting in the following way. Let $S_1, \ldots, S_n$ denote independent latent factors underlying the variation of the data, and let them follow some prior distribution $p(S) = \prod_i p(S_i)$. We assume that they produce the observations $X$ via a deterministic function $g$, i.e., $X = g(S)$, yielding a distribution $p(X)$, and that $S$ can be recovered via an encoder $S = e(X)$. Unfortunately, our training set suffers from some bias, giving rise to unwanted correlations between the factors of variation as expressed by some other distribution $p'(S)$. We merely assume that it is faithful to the true distribution on the level of single variable marginals, i.e., $p'(S_i) = p(S_i)$, which is typically relatively easy to ensure. We call the resulting biased distribution of our training data $p'(X)$.

Our goal is to obtain a generative model of $p(X)$ based on samples from $p'(X)$ in combination with small number of examples providing weak supervision. We follow the VAE framework, introducing a latent code $Z = Z_1, \ldots, Z_m$ with $m \geq n$, and fix an independent prior distribution over the $Z_i$, e.g., a unit Gaussian. Via a learned decoder $\hat{g}(Z)$, we would like to generate samples matching $p(X)$, i.e. $p(\hat{g}(Z)) \approx p(X)$.

Suppose that we have managed to perfectly disentangle the VAE’s latent codes with regard to the true factors of variation, by ensuring that each $Z_i$ captures exactly $S_i$ for $i \leq n$. Then there exist bijections $f_i$ deterministically mapping the two to each other: $Z_i = f_i(S_i)$, and their marginal probabilities will match, i.e., $p(S_i) = p'(S_i) = p(f(S_i)) \frac{d}{dS_i} f(S_i)$.

If in addition, the VAE’s decoder $\hat{g}$ approximately matches the data generating procedure such that $\hat{g}(f(S)) \approx g(S)$, then it follows that the distribution $\hat{p}(X)$ represented by the VAE matches the desired unbiased distribution $p(X)$. Letting $\delta_X[\cdot]$ denote the Dirac delta at $X$, we have

$$p(X) = \int \delta_X(g(S)) \prod_i p(S_i) dS$$

$$\approx \int \delta_X(\hat{g}(f(S))) \prod_i p(f(S_i)) \left| \frac{d}{dS_i} f(S_i) \right| dS$$

$$= \int \delta_X(\hat{g}(Z)) \prod_i p(Z_i) dZ = \hat{p}(X),$$

where the last line follows from substituting $Z = f(S)$, cancelling out the differentials.

Figure 2: Our VAE uses a small amount of human feedback in addition to the train set to learn disentangled representations. For the sake of simplicity, only augmented loss terms are illustrated here.
restricted to where \( L \) to prove that the additional label-based supervision we introduce is sufficient for learning restrictive and non-trivial nuisance factors. Since our method builds on match pairing, its guarantees trivially transfer. Therefore, we only need two cases: Disentangling all factors, and disentangling only a subset of factors, leaving the rest as arbitrarily encoded.

Table 1: Properties of our and match pairing approach in case of disentangling a subset of factors (shortly denoted as “subset” in the Tab.) and disentangling all factors (shortly denoted as “all factors” in the Tab.)

|                | Consistency | Restrictiveness | DISENTANGLEMENT (Consistency & Restrictiveness) | Non-Trivial Guarantee |
|----------------|-------------|-----------------|-----------------------------------------------|-----------------------|
| Ours           | ✓           | ✓               | ✓                                             | ✓                     |
| Match          | ✓           | x               | ✓                                             | x                     |

3.1 Disentanglement through Human Feedback

Based on this insight, we design a combination of loss functions to encourage the model to fit the training data as well as possible while enforcing disentanglement. The former is achieved using the standard ELBO loss. To ensure that a latent variable \( Z_i \) only contains information pertaining to target factor \( S_i \), we collect pairs of samples \( X, X' \) which share the same value for \( S_i \), while the other factors \( S_{\setminus i} \) vary at random. Following Shu et al. [2020], we minimize the match-pairing loss \( L_{MP}^i(X, X') = \mathbb{E}[|Z_i - Z'_i|^2] \) for those samples, where \( Z_i, Z'_i \) are sampled from \( q(Z_i|X), q(Z_i|X') \). In addition, we use a small number of samples \( X \) for which the true value for the target underlying factor \( S_i \) is available. To ensure that \( Z_i \) is predictive for \( S_i \), while the rest of the latent variables \( Z_{\setminus i} \) is not, we train regression models in parallel with the main model to predict \( S_i \) from \( Z_i \) and \( Z_{\setminus i} \), respectively. We then use these models to compute the classification loss \( L_{CL}^i(X) = \mathbb{E}_{q(Z_i|X)}[\lambda_2 R_i(S_i, Z_i) - \lambda_3 R'_i(S_i, Z_{\setminus i})] \), where \( R_i \) and \( R'_i \) denote the cross-entropy classification loss of the two classifiers. The way these losses interlock is illustrated in Fig. 3. The full loss for our disentangled VAE is \( L(X) = -\text{ELBO}(X) + \sum_i \lambda_1 L_{MP}^i(X) + L_{CL}^i(X) \),

\[
(4)
\]

where \( L_{MP}^i \) and \( L_{CL}^i \) is only computed for all datapoints and factors of variation \( i \) for which the required augmentations are available. Fig. 2 illustrates the resulting computation graph.

3.2 Analysis of Theoretical Guarantees

Following the arguments from Locatello et al. [2019b], we prove that disentanglement arises from our supervision method instead of model inductive bias. We use the definition of disentanglement from Shu et al. [2020], where disentanglement is decomposed into consistency and restrictiveness. Strong disentanglement is guaranteed when both consistency and restrictiveness hold. Let \( \hat{e}_i \) denote a learned deterministic encoder mapping from \( X \) to \( Z_i \). According to the definition of Shu et al. [2020], \( S_i \) is consistent with \( Z_i \) if \( E[||\hat{e}_i \circ g(S_i, Z_{\setminus i}) - \hat{e}_i \circ g(S_i, Z_{\setminus i})||^2] = 0 \), and \( S_i \) is restricted to \( Z_i \) if \( E[||\hat{e}_i \circ g(S_i, Z_{\setminus i}) - \hat{e}_i \circ g(S_i', Z_{\setminus i})||^2] = 0 \). Intuitively, consistency expresses that \( Z_i \) does not capture the factors of variation other than \( S_i \), and restrictiveness expresses that \( S_i \) is only captured in \( Z_i \). In addition, Shu et al. [2020] discuss the property that \( Z \) should be non-trivial, i.e., it should encode \( S \) at all.

Tab. 1 summarizes the theoretical guarantees of each property of our method and pure match pairing. We distinguish two cases: Disentangling all factors, and disentangling only a subset of factors, leaving the rest as arbitrarily encoded nuisance factors. Since our method builds on match pairing, its guarantees trivially transfer. Therefore, we only need to prove that the additional label-based supervision we introduce is sufficient for learning restrictive and non-trivial solutions when disentangling a subset of factors.
Table 2: Summary of the datasets. Parentheses give number of quantized values for each target factor. The last column gives the size of samples we chose from the train set to perform data augmentation on.

| Name               | Training Instances | Held-out Instances | Image Size   | Ground-truth Factor          | Data Augmentation |
|--------------------|--------------------|--------------------|--------------|-----------------------------|-------------------|
| COLORED MNIST      | 60K                | 10K                | $28 \times 28 \times 3$ | SHAPE(10), COLOR(10)       | 600               |
| COLORED dSprites   | 10K                | 10K                | $64 \times 64 \times 3$ | SHAPE(3), COLOR(3), NUISANCE | 1K                |
| 3D SHAPES          | 10K                | 10K                | $64 \times 64 \times 3$ | OBJECT SHAPE(4), OBJECT COLOR(4), NUISANCE | 1K                |

Define a hypothesis space $\mathcal{H}$ of models we are willing to consider. Let $(p(S), p'(S), g, e) \in \mathcal{H}$ denote an arbitrary ground-truth model which generated the data we are observing. Following [Shu et al., 2020], we call a supervision method $S : \mathcal{H} \rightarrow \mathcal{P}$ sufficient for a certain guarantee if there exists a learning algorithm $\mathcal{A} : \mathcal{P} \rightarrow \mathcal{H}$ which uses the supplied examples to produce a learned model $(p(Z), \hat{g}, \hat{e})$ for which the desired guarantee holds. In the following, we show that our method of supervision is sufficient for non-trivial and restrictive solutions. We do so by considering an $\mathcal{A}^*$ which obtains the global minimum of the classification loss $L_{CL}^*$, despite using universal approximators for $\hat{R}$ and $\hat{R}'$.

**Non-Triviality Guarantee.** We claim that our method guarantees the learned model $(p(z), \hat{g}, \hat{e})$ provides non-trivial solutions for $Z_i$, i.e. there is a learning algorithm that for which $S_i \perp \perp Z_i$. Match pairing alone does not entail this guarantee when disentangling only a subset of factors. To see why, consider a simple counterexample with underlying factors $S_1, S_2 \sim \mathcal{N}(0, 1)$, and generator $g(S) = S_1 \cdot S_2$. There exists a learning algorithm $\mathcal{A}$ which optimizes the match pairing loss on factor $S_1$ by setting $(g(S_1, S_2), g(S_1, S_2')) = (\hat{g}(Z_1, Z_2), \hat{g}(Z_1, Z_2'))$, and $\hat{g}(Z) = [Z_1, Z_2] = [0, Z_1' \times Z_2']$ where $Z_1' \sim \mathcal{N}(0, 1)$, $Z_2' \sim \mathcal{N}(0, 1)$. In this case, match pairing yields a trivial solution for $Z_1$, i.e. $Z_1 = 0$.

**Proof:** A learned model $(p(z), \hat{g}, \hat{e})$ provided by $\mathcal{A}^*$ minimizes the classification loss, and the positive term $\hat{R}'(Z_i, S_i)$ in particular. Consequently, it is possible to accurately predict $S_i$ from $Z_i$, and $S_i \perp \perp Z_i$. ■

**Restrictiveness Guarantee.** We postulate that our method guarantees that the learned model $(\hat{p}(z), \hat{g}, \hat{e})$ restricts the information of $S_i$ to $Z_i$. That is, $Z_i \perp \perp S_i$.

**Proof:** Suppose to the contrary that $Z_i \perp \perp S_i$. Then we could find a regression model that predicts $S_i$ better than random, i.e., the value of $-\hat{R}'(Z_i, S_i)$ would be suboptimal. This is a contradiction, since we assumed that $\mathcal{A}^*$ would achieve the global minimum of $L_{CL}^*$. ■

4 Empirical Investigation

Our main intention here is to address the following question empirically: *Does our proposed model disentangle the latent space better than the mainstream counterpart models?* To this end, we ran a series of experiments for both qualitative and quantitative evaluation. For the qualitative comparison, we considered reconstructed images and latent space arithmetic for generating novel content, as well as latent space traversal.

**Dataset.** We considered several standard benchmark datasets. First, we generated colored MNIST [LeCun et al., 2010] in a similar way as [Kim et al., 2019]. In this domain, we assumed two latent factors of variations: shape and color. In train set, every shape class of images was randomly and consistently assigned with a color so the shape class and the color class correlated with each other. In test set, one could assign random colors to each image. But we made the problem even more difficult by assigning the color class reversely correlated to the shape class so that none of these combinations of shape and color were seen in the train set. As another dataset, we generated colored dSprites [Higgins et al., 2016] in a similar spirit. We randomly sampled 10k instances as train set and injected confounding in it by consistently assigning one color to each shape category and the rest factors remained untouched. In test set, we assigned each color to the shape with one offset as to in the train set, so that none of these two factors’ combinations were seen in the train set. The third dataset is 3D Shapes [Kim and Mnih, 2018]. We sampled 10k instances to be the train set whereby each object hue category correlated to one object shape category. In test set, each object hue category correlated to the object shape category with one offset as to in the train set for the same reason.
Figure 4: Reconstruction images from different models on colored MNIST, colored dSprites and 3D shapes. Each shape factor is consistently associated with one color factor in the train set, while in the test set the association is different. Some images from the baseline models are reconstructed wrongly with the shape (color) factor that associated with the corresponding color (shape) factor in the train set, implying entangled latent representations.

Figure 5: Latent space traversal of our VAE. The first (last) 4 rows are respectively the traversal of $Z_0$, $Z_1$, $Z_2$ and $Z_3$ ($Z_4$, $Z_5$, $Z_6$ and $Z_7$) which are collectively intended for shape (color) factor. The samples show that varying the first (last) 4 variables indeed leads to varying object shape (color), which suggests that each unit of representation has captured the intended factor. Meanwhile, the nuisance factors such as wall color stay unchanged, which suggests that the nuisance factors are not captured in these 8 variables.

Data Augmentation to Emulate a Human User. We augmented the biased dataset with a few amount of examples to emulate human feedback. Specifically, on each target ground-truth factor $S_i$, we generate a few samples $X'_i$ that vary on the rest factors $S_{\setminus i}$ while keeping the target factor fixed, and a few samples $X''_i$ that vary only on the target factor while keeping the rest factors fixed. That is, $X'_i$ share the target factor and only the target factor, and $X''_i$ share the rest factors and only the rest factors. In addition, we acquired a few $S_i$ of $X'_i$ to avoid trivial solutions of $Z_i$ (non-trivial guarantee) by enforcing $Z_i$ to contain information about $S_i$, and we constrained that the nuisance variables $Z_{\setminus i}$ do not contain information about $S_i$ (restrictiveness guarantee).
Figure 6: Accuracies of the post-hoc linear classifiers after debiasing for shape (top row) and color (bottom row). 1: $\beta$-VAE, 2: $\beta$-TCVAE, 3: FactorVAE, 4: Ada-GVVAE, 5: Our VAE without annotations, 6: Our VAE with annotations. The higher the value, the better the performance.

Figure 7: Several disentanglement metrics (FactorVAE score, MIG, DCI disentanglement and DCI completeness) across different models on colored MNIST (top row), colored dSprites (middle row) and 3D shapes (bottom row). 1: $\beta$-VAE, 2: $\beta$-TCVAE, 3: FactorVAE, 4: Ada-GVVAE, 5: Our VAE without annotations, 6: Our VAE with annotations. The higher the value, the better the performance.

Experimental Protocol. We followed the experimental protocol of Locatello et al. [2019c, 2020] and aligned with the general literatures on disentanglement [Higgins et al. [2016], Kim and Mnih [2018], Chen et al. [2018]]. The baseline models used are $\beta$-VAE [Higgins et al. [2016], FactorVAE [Kim and Mnih [2018]], $\beta$-TCVAE [Chen et al. [2018]] and Ada-GVVAE [Locatello et al. [2020]]. These models are mainstream VAEs for benchmarking disentangled representations. In addition, we also evaluated our VAE with unannotated human feedback to investigate the value of the annotations. We implemented all models in python and pytorch. For each model, we ran experiments with 8 random seeds. For each baseline model, we used 3 fixed hyperparameters on each dataset. Our model’s hyperparameters were chosen empirically according to the quality of the reconstructions. Our experience suggests using hyperparameters $\lambda_1 = \lambda_2$ and $\lambda_3 = 1$ or $\lambda_3 = 10$ lead to good results in practice, so we propose to tune the hyperparameters within this
range. We used linear classifiers for $R_s$ and $R_t'$ in Eq. 4. We relaxed the condition for $R_t'$ to be an universal classifier in practice and demonstrated that this can still yield promising disentangled representations empirically.

4.1 Qualitative Evaluation

To gain a qualitative and intuitive insights, let us start off by investigating reconstructed images.

**Reconstructions.** As mentioned above, we chose the test samples to be completely new to the trained models by sampling on the latent space where the combinations of the target latent factors were never seen in the train set. In this case, if the target latent factors were entangled in the learned representations, the model would struggle to encode all the target latent factors in this unseen combination, this in turn would lead to unsatisfactory reconstructions. Fig. 4 shows some randomly chosen test samples across the three datasets and their reconstructions via different models. One can see for colored MNIST and colored dSprites, the reconstructed color via the baseline models for some samples still exhibited the color that associated with those samples’ shape in the train set. However, this happens much less often on our model. On 3D shapes, the object color is mostly reconstructed reasonably, but for the baseline models, the reconstructed object shapes often still exhibit the shape that associated with the corresponding object color in the train set, implying for entangled latent representations. For our model, both the object shape and the object color were reconstructed very plausibly. The model achieved this robustness w.r.t. generalization on the unseen samples by disentangling the latent representations.

**Latent Space Arithmetic.** This experiment demonstrates that our model can not only generalize under non-i.i.d distributions, but can also generate novel content. Figs. 1(right) and 8 show all possible hybridization of shapes and colors generated by our VAE in each corresponding domain. Each hybridized image was generated by concatenating the latent representation of one image’s shape on the leftmost column and another image’s color on the first row. When nuisance variables are present, they take the values of the average of both reference images’ nuisance representations. The result shows that the hybridized images very plausibly resemble both the shape from the leftmost-column images and the color from the first-row images, and this yields and recovers the whole spectrum of the target domain where only a small fraction of which were seen in the train set. This not only allows us to generate novel contents, but also grants us opportunities to manipulate the latent space and generate specific samples for downstream tasks.

**Latent Space Traversal.** To give a better overview of the latent space and what the generative models have really learnt for the target factors, we show partial latent space traversals of our model for 3D shapes. The latent variables are in total 50. We designated the first 4 variables for learning the shape factor, and the second 4 variables for the color factor, the rest are considered nuisance variables. Fig. 5 shows samples generated by traversing the first 8 latent variables. The first (last) 4 rows correspond to the variables that are supposed to learn the shape (color) factor. The samples suggest that each unit of the representations has reasonably learnt the target factor and is also mostly restricted to that factor. Moreover, the nuisance factors, wall color etc, seem not to be captured by these 8 variables which implies the latent representations are consistent with the target factors.

4.2 Quantitative Evaluation

For a quantitative evaluation of the disentanglement, we used FactorVAE score Kim and Mnih [2018], DCI Disentanglement and DCI Completeness Eastwood and Williams [2018] and Mutual Information Gap (MIG) Chen et al. [2018]. As these metrics are not conceptualized accounting for biased data, we use the whole spectrum of the target domain to evaluate them.

DCI Disentanglement measures the degree to which each latent variable capturing at most one generative factor. This score gives a direct hint on the consistency property (see Tab. 1). DCI Completeness measures the degree to which each factor is captured by a single latent variable. Although our model does not assume single latent variable for each factor and therefore this metric is not in favor of us, this score still gives a hint on how densely the information of $S_i$ is packed in the latent variables, and in turn on the restrictiveness property (see Tab. 1).

Since MIG enforces axis-alignment and our VAE does not make this assumption, this metric in its original form is not fit for evaluating our VAE. Therefore we adapt MIG by constraining the gap computation between the known unit of the latent representations and the rest representations. In particular, we slightly reformulate it to the following form:

\[
\frac{1}{n} \sum_{i=1}^{n} \frac{1}{H(S_i)} \left( \max I(Z_i; S_i) - \max I(Z_{\backslash i}; S_i) \right)
\]

where $I(Z_i; S_i)$ denotes the mutual information between the ground-truth factor $S_i$ and the unit of the latent representations designated for $S_i$, i.e. $Z_i$. $I(Z_{\backslash i}; S_i)$ denotes the mutual information between $S_i$ and the rest of the latent representations $Z_{\backslash i}$. The higher this adapted MIG is, the information about $S_i$ is more densely packed in $Z_i$. Since the baseline models make the axis-alignment assumption, we use the original form of MIG to evaluate them.
Fig. 7 shows the distribution of these scores from different models across different datasets via violin graphs. All the metrics are bounded by 0 and 1. Higher values imply better performance. As one can see, our method yields significantly better performance in terms of all the metrics. In terms of FactorVAE score, although the baseline models have also achieved almost the upper bound of FactorVAE score by its best performance, our model still has less variance across different runs, especially on colored MNIST we reached 1 for every random run. This demonstrates that our method is significantly better at learning disentangled representation on the considered datasets.

**Downstream Accuracy.** For each VAE, we trained a post-hoc linear classifier on each unit of the latent representations to predict the values of the corresponding factor. We use the original train set (without augmented data) where each shape is consistently associated with one color so that we can fairly compare the quality of the learned representations and exclude the influence of differently distributed train data for each classifier. For evaluation, we generated data where shape and color have different associations. This way, we make the test set a different distribution to the train set on purpose, so we can compare whose representations allow better generalization and robustness under covariate shifts Quionero-Candela et al. [2009].

Fig. 6 reports the distribution of the test set accuracies. Apparently our model results in significantly better downstream accuracies, achieving 100% accuracies at its best. This is a compelling evidence that our model yields more robust features for the downstream classifiers on the considered datasets.

**5 Conclusion and Future Work**

We presented a method for learning unbiased generative models using a biased training set which violates the i.i.d. assumption. It obtains disentangled representations by incorporating additional human feedback in the form of a small amount of examples with partially available label information. Our empirical study confirmed the effectiveness of our approach at debiasing the data distribution, which in turn yields better generalization to unbiased test sets.

Our approach could benefit many real world tasks which should be investigated in future work. One particular focus should be to make the learning process fully interactive. In addition, one could work on generalizing this approach to other types of generative models.

**References**

Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1521–1528. IEEE, 2011.

Tatiana Tommasi, Novi Patricia, Barbara Caputo, and Tinne Tuytelaars. A deeper look at dataset bias. In Domain adaptation in computer vision applications, pages 37–55. Springer, 2017.
Samuel Ritter, David GT Barrett, Adam Santoro, and Matt M Botvinick. Cognitive psychology for deep neural networks: A shape bias case study. In Proceedings of International conference on machine learning, pages 2940–2949. PMLR, 2017.

Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In Proceedings of the European Conference on Computer Vision (ECCV), pages 498–512, 2018.

Aditya Grover, Kristy Choi, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In Proceedings of International Conference on Machine learning, 2020.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. Nature communications, 10(1):1–8, 2019.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In Proceedings of Advances in Neural Information Processing Systems, 2016.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In Proceedings of International Joint Conference on Artificial Intelligence, pages 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL: https://doi.org/10.24963/ijcai.2017/371.

W James Murdoch, Peter J Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from lstms. arXiv preprint arXiv:1801.05453, 2018.

Stefano Teso and Kristian Kersting. Explanatory interactive machine learning. In Proceedings of the 2nd AAAI/ACM Conference on AI, Ethics, and Society, AIES, 2019.

Laura Rieger, Chandan Singh, W James Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. arXiv preprint arXiv:1909.13584, 2019.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Learning explainable models using attribution priors. arXiv preprint arXiv:1906.10670, 2019.

Ramprasaath R Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a hint: Leveraging explanations to make vision and language models more grounded. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2591–2600, 2019.

Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Franziska Herbert, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence, 2(8):476–486, 2020.

Xiaoting Shao, Arseny Skryagin, Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for better reasons: Training differentiable models by constraining their influence function. In Proceedings of AAAI Conference on Artificial Intelligence, 2021.

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.

Aditya Grover, Jiaming Song, Alekh Agarwal, Kenneth Tran, Ashish Kapoor, Eric Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. In Proceedings of Advances in Neural Information Processing Systems, 2019.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In Proceedings of International Conference on Learning Representations, 2019.

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In Proceedings of International Conference on Machine Learning, 2020.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In Advances in Neural Information Processing Systems, pages 14245–14258, 2019.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In Proceedings of Advances in Neural Information Processing Systems, pages 14611–14624, 2019a.
Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. 2019.

Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David E Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Disentangled representation learning in cardiac image analysis. Medical image analysis, 58:101535, 2019.

Irina Higgins, Arka Pal, Andrei A Rusu, Loic Matthey, Christopher P Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. 2017.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. CoRR, abs/1312.6114, 2014.

Junxiang Chen and Kayhan Batmanghelich. Weakly supervised disentanglement by pairwise similarities. Proceedings of the AAAI Conference on Artificial Intelligence, 2020.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. Proceedings of the 37th International Conference on Machine Learning, 2020.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. Proceedings of the International Conference on Learning Representations, 2020.

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.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. arXiv preprint arXiv:1711.00848, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. 2018.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In Proceedings of Advances in Neural Information Processing Systems, pages 2610–2620, 2018.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. arXiv preprint arXiv:1804.03599, 2018.

Emile Mathieu, Tom Rainforth, N Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In International Conference on Machine Learning, pages 4402–4412, 2019.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In Proceedings of International Conference on Machine learning, pages 4114–4124, 2019b.

Karl Ridgeway and Michael C Mozer. Learning deep disentangled embeddings with the f-statistic loss. In Proceedings of Advances in Neural Information Processing Systems, pages 185–194, 2018.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In Proceedings of AAAI Conference on Artificial Intelligence, 2018.

Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In Proceedings of International Joint Conference on Artificial Intelligence, pages 2506–2513, 2019.

Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variation using few labels. In Proceedings of International Conference on Learning Representations, 2019c.

Peter Sorrenson, Carsten Rother, and Ullrich Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). arXiv preprint arXiv:2001.04872, 2020.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In Proceedings of International Conference on Artificial Intelligence and Statistics, pages 2207–2217, 2020.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2, 2010.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In Proceedings of International Conference on Learning Representations, 2018.

Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. Dataset shift in machine learning. The MIT Press, 2009.