DISENTANGLEMENT AND GENERALIZATION UNDER CORRELATION SHIFTS

Christina M. Funke*
University of Tübingen

Paul Vicol*
University of Toronto
Vector Institute

Kuan-Chieh Wang
University of Toronto
Vector Institute

Matthias Kümmerer†
University of Tübingen

Richard Zemel†
University of Toronto
Vector Institute

Matthias Bethge†
University of Tübingen

ABSTRACT

Correlations between factors of variation are prevalent in real-world data. Exploiting such correlations may increase predictive performance on noisy data; however, often correlations are not robust (e.g., they may change between domains, datasets, or applications) and models that exploit them do not generalize when correlations shift. Disentanglement methods aim to learn representations which capture different factors of variation in latent subspaces. A common approach involves minimizing the mutual information between latent subspaces, such that each encodes a single underlying attribute. However, this fails when attributes are correlated. We solve this problem by enforcing independence between subspaces conditioned on the available attributes, which allows us to remove only dependencies that are not due to the correlation structure present in the training data. We achieve this via an adversarial approach to minimize the conditional mutual information (CMI) between subspaces with respect to categorical variables. We first show theoretically that CMI minimization is a good objective for robust disentanglement on linear problems. We then apply our method on real-world datasets based on MNIST and CelebA, and show that it yields models that are disentangled and robust under correlation shift, including in weakly supervised settings.

1 INTRODUCTION

Disentangled representations can be useful for improving fairness (Locatello et al., 2019a), interpretability (Adel et al., 2018), controllable generative modeling (He et al., 2019), and transfer to downstream tasks (Van Steenkiste et al., 2019). In addition, they can improve robustness on out-of-distribution data (Higgins et al., 2017b) (e.g., for domain adaptation (Ilse et al., 2020) and domain generalization (Ben-Tal et al., 2009)). Most research on disentanglement has assumed that the underlying factors of variation in the data are *independent* (e.g., that factors are not correlated). However, this assumption is often violated in real-world settings: for example, in domain adaptation, the class distribution often shifts between domains (yielding a correlation between the class and domain); in natural images, there is often a strong correlation between the foreground and background (Beery et al., 2018), or between multiple foreground objects that tend to co-occur (e.g., a keyboard and monitor) (Tsipras et al., 2020; Beyer et al., 2020). Importantly, correlated data occur in areas that affect people's lives, including in healthcare (Chartsias et al., 2018) and fairness applications (Madras et al., 2018; Creager et al., 2019; Locatello et al., 2019a), and correlation shifts in these applications are common (e.g., demographics are likely to differ from one hospital to another).

The goal of disentanglement is to encode data into independent subspaces that preferably match the ground truth generative factors. A common approach to achieve this (used in ICA, PCA, and VAEs) is to ensure that the latent subspaces share as little information as possible, by minimizing the mutual information (MI) between subspaces. However, recently it has been shown that this fails to disentangle correlated factors (Träuble et al., 2020). Several works have sought to address this by introducing partial supervision (Träuble et al., 2020; Shu et al., 2019; Locatello et al., 2020b). Here, we show that even with *full* supervision, minimizing the MI can fail: it is impossible to encode generative factors into independent subspaces if they are correlated in the training data. To address this, we propose minimizing the MI between subspaces *conditioned* on the correlated attributes.

We compare three objective functions for learning disentangled representations: 1) standard supervised losses (such as mean-squared error or cross-entropy) that encourage each subspace to encode a specific attribute; 2) a supervised

* Equal contribution. † Shared senior authors.

loss plus *unconditional* MI minimization; and 3) a supervised loss plus *conditional* MI (CMI) minimization. We first show that approaches (1) and (2) fail on correlated and noisy data: minimizing a supervised loss cannot enforce that there is little information shared between subspaces; MI minimization is too strong a constraint to satisfy when the underlying factors of variation are correlated, and thus minimizing MI leads to decreased performance. We then show that minimizing CMI yields disentangled representations that are robust to correlation shifts.

Overall, we aim to establish conditional independence as the correct notion of independence between latent subspaces when disentangling data with correlated factors of variation.

Contributions.

- Most disentanglement metrics used in the literature assume that the attributes are uncorrelated, and thus are not directly applicable to correlated data. We propose to use the *predictive performance under correlation shift* as a *measure of disentanglement* applicable to settings with correlated factors of variation.
- We analyze the behavior of each objective function on a linear regression problem where all quantities of interest can be computed analytically (Section 3). We show that minimizing the CMI between latent subspaces yields a solution robust to test-time correlation shifts, while minimizing the unconditional MI (or only a supervised loss) does not.
- We describe an adversarial approach for learning conditionally disentangled representations (Section 4).
- Then, we apply our approach to CMI minimization to two tasks based on real-world datasets—a multi-digit occluded MNIST task and correlated CelebA—and demonstrate improved performance under correlation shift relative to baselines (Section 5).
- We investigate the interplay between correlation strength and noise level in the training data. When data are noisy and have strong correlations, the noise forces the model to rely on correlations when making a prediction; this leads to failures of the baseline approaches when correlations shift at test-time, and demonstrates the benefits of CMI minimization, which performs well across correlation strengths and noise levels.
- We show that CMI minimization can be applied in the weakly supervised setting, and show significant gains compared to baselines.

Our code is available on Github.

2 Background & Related Work

ICA/ISA. Disentanglement is related to blind source separation (BSS), as both problems revolve around the question of identifiability. A classic approach to BSS is Independent Component Analysis (ICA) (Comon, 1994; Jutten & Herault, 1991; Bell & Sejnowski, 1997; Olshausen & Field, 1996), which assumes statistical independence between the source variables (Jutten & Herault, 1991; Jutten & Karhunen, 2003). Independent Subspace Analysis (ISA) (Hyvärinen & Hoyer, 2000), or multidimensional ICA (Cardoso, 1998), is a generalization of ICA where each component is a k-dimensional subspace; dimensions within a subspace may have dependencies, while dimensions from different subspaces must be independent. Our work can be seen as a form of nonlinear ISA that enforces conditional independence between subspaces.

Correlations Between Features. With roots in ICA, most research on disentanglement focuses on data that was generated by independent factors, including synthetic benchmarks such as dSprites (Matthey et al., 2017), Shapes3D (Burgess & Kim, 2018), Cars3D (Reed et al., 2015), SmallNORB (LeCun et al., 2004), or MPI3D (Gondal et al., 2019). In real-world datasets on the other hand, factors are often correlated (Welinder et al., 2010; Lin et al., 2014). Träuble et al. (2020) pointed out the challenges that arise when attempting to learn disentangled representations on correlated data, and performed a large-scale empirical evaluation of the effect of correlations on widely-used VAE-based disentanglement models. They proposed two approaches to ameliorate the harmful effects of correlations: 1) introducing weak supervision during training, and 2) labeling data post-hoc to "correct" a pre-trained encoder. We show that even with full supervision, correlations are problematic when enforcing independence between latent subspaces. Causally-informed modeling (Zhang et al., 2020) is another approach to learning disentangled representations and extracting invariant features.To investigate the effect of correlations systematically, it is common to modify existing datasets to induce correlations, for example by subsampling the data, or generating synthetic datasets with the desired properties (Dittadi et al., 2020; Cimpoi et al., 2014; Jacobsen et al., 2018; Locatello et al., 2019b). We follow this approach in our experiments.

Unsupervised and Weakly-Supervised Disentanglement. Disentangled representation learning is often studied in the unsupervised setting, where the ground-truth factors of variation are unknown. Widely-used approaches for this include variational autoencoders (VAEs) (Kingma & Welling, 2013) and their variants (beta-VAE (Higgins

et al., 2017a), TC-beta-VAE (Chen et al., 2018), FactorVAE (Kim & Mnih, 2018), etc.). However, it was shown by Locatello et al. (2019b) that the assumption of independent source variables (e.g., attributes) is questionable, and that *purely unsupervised* disentanglement may not be possible. This spurred interest in *weakly-supervised* methods (Shu et al., 2019; Locatello et al., 2020b), where weak supervision is provided in the form of partial labels or grouping information (Bouchacourt et al., 2018; Nemeth, 2020; Klindt et al., 2020). In this paper, we focus on comparing MI and CMI minimization in the fully-supervised setting, as this is already challenging and provides useful insights.

Domain Adaptation/Generalization. We use predictive performance under correlation shift as a measure for the quality of disentanglement. This is closely related to the fields of domain adaptation and generalization, with the difference that we assume access to one source domain only. The goal of most related work in this field is to learn representations from multiple source domains that transfer to known (e.g., adaptation) or previously unseen (e.g., generalization) target domains. This is done by either learning domain-invariant representations which discard domain information (Tzeng et al., 2017) or by learning disentangled representations, with latent subspaces that correspond to the domain and the class, respectively (Peng et al., 2019; Ilse et al., 2020; Liu et al., 2018). For the latter approach, disentanglement is achieved by minimizing the mutual information between latent subspaces (Cheng et al., 2020; Gholami et al., 2020; Nemeth, 2020). Zhao et al. (2019) discuss fundamental problems inherent in learning domain-invariant representations when there are correlations between classes and domains (e.g., when the class distribution shifts in the target domain). The goal of Invariant Risk Minimization (Arjovsky et al., 2019) is to find correlations that are invariant over multiple training domains in order to improve generalization to out-of-distribution data.

Fairness. An important application of disentanglement is fairness. As machine learning systems are typically trained on historical data, they often inherit past biases (e.g., from human decision-makers). This may result in unfair treatment on the basis of sensitive properties such as ethnicity, gender, or disability. Typically, this can be addressed by modifying the training data to be unbiased or by adding a regularizer (e.g. based on mutual information) that quantifies and minimizes the degree of bias (Kamiran & Calders, 2009; Kamishima et al., 2011; Zemel et al., 2013; Hardt et al., 2016; Cho et al., 2020).

Mutual Information. The mutual information (MI) between two random variables $\mathbf{x}$ and $\mathbf{y}$, denoted $I(\mathbf{x};\mathbf{y})$, is the KL divergence between the joint distribution $p(\mathbf{x},\mathbf{y})$ and the product of the marginal distributions $p(\mathbf{x})p(\mathbf{y})$: $I(\mathbf{x};\mathbf{y}) = D_{\mathrm{KL}}[p(\mathbf{x},\mathbf{y})||p(\mathbf{x})p(\mathbf{y})]$. Minimization of MI has been used to implement an information bottleneck (Alemi et al., 2016) and to factorize representations (Jacobsen et al., 2018). MI minimization is at the heart of many approaches to disentanglement. The *conditional mutual information* (CMI) is defined as: $I(\mathbf{x};\mathbf{y} \mid \mathbf{z}) = \mathbb{E}_{\mathbf{z}}\left[D_{\mathrm{KL}}[p(\mathbf{x},\mathbf{y} \mid \mathbf{z}) \mid\mid p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})\right]$. CMI measures the dependency between two variables given that we know the value of a third variable. For example, there is a dependency between a country's number of Nobel laureates per capita and chocolate consumption per capita (Prinz, 2020). However, this dependency is largely explained by the wealth of a country, thus $I(nobel;chocolate \mid wealth) < I(nobel;chocolate)$. In general, the CMI can be smaller or larger than the unconditional MI.

Estimating & Optimizing Mutual Information. Many approaches have been proposed for MI and CMI estimation and optimization. The Mutual Information Neural Estimator (MINE) (Belghazi et al., 2018) uses a lower-bound of the MI based on the Donsker-Varadhan dual representation of the KL divergence (Donsker & Varadhan, 1983). Poole et al. (2019) provide an overview of variational bounds that can be used to estimate MI; most are *lower bounds*, which are useful in principle for *maximizing* MI, but which have also been used to minimize MI (even though minimizing a lower bound is not guaranteed to decrease MI). CLUB (Cheng et al., 2020) introduced a variational upper bound of MI, providing a more principled objective for minimizing MI. Several CMI estimators have been proposed, including conditional-MINE (Molavipour et al., 2020a), C-MI-GAN (Mondal et al., 2020), CCMI (Mukherjee et al., 2020), and an approach based on nearest neighbors (Molavipour et al., 2020b). Many approaches to MI minimization are based on batchwise shuffling of latent subspaces, sometimes referred to as metameric sampling (Belghazi et al., 2018; Nemeth, 2020; Feng et al., 2018; Park et al., 2020; Peng et al., 2019). The approach we use in Section 4 follows this paradigm of latent-space shuffling.

3 Disentanglement with Correlated Variables: Motivating CMI

A summary of notation is provided in Appendix A.

Problem Statement. Suppose we observe noisy data $\mathbf{x} \in \mathbb{R}^m$ obtained from an (unknown) generative process $\mathbf{x} = g(\mathbf{s})$ where $\mathbf{s} = (s_1, s_2, \ldots, s_K)$ are the *underlying factors of variation*, also called source variables or attributes, which may be correlated with each other. We wish to find a mapping $f : \mathbb{R}^m \to \mathbb{R}^n$ to a latent space $f(\mathbf{x}) = \mathbf{z} =$

	Base	**Base + MI**	**Base + CMI**
Variance Explained, Training (Corr = 0.8)	91.9%	69.8%	90.9%
Variance Explained, Test (Corr = 0)	87.6%	65.0%	90.9%
Regression Matrix M (where $\hat{\mathbf{s}} = M\mathbf{x}$)	$\begin{pmatrix} 0.81 & 0.14 \\ 0.14 & 0.81 \end{pmatrix}$	$\begin{pmatrix} 1.07 & -0.46 \\ -0.46 & 1.07 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

Table 1: Robustness of linear regression under correlation shift for each of the objectives *Base*, *Base+MI*, and *Base+CMI*. Here, the observations and predictions are in $\mathbb{R}^2$. The performance of the *Base* model drops under correlation shift. The optimal solution under the constraint of minimal MI, $I(z_1; z_2) = 0$, fails to model the in-distribution correlated training data. The solution with minimal *conditional* MI, $I(z_1; z_2 \mid s_1) = I(z_1; z_2 \mid s_2) = 0$, maintains consistent performance under correlation shift. Note that because the generative process is given by $g(\mathbf{s}) = \mathbf{A}\mathbf{s} = \mathbf{I}\mathbf{s}$, the inverse is $\mathbf{A}^{-1} = \mathbf{I}$. In the last row, we see that only Base + CMI recovers this true inverse.

$(\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_K)$ such that each attribute s_k can be recovered from the corresponding latent subspace $\mathbf{z}_k$ by a linear mapping $\mathbf{R}_k$, e.g., $\hat{s}_k = \mathbf{R}_k \mathbf{z}_k$ such that $\hat{s}_k \approx s_k$. We denote by $\mathbf{z}_{-i}$ the set of subspaces $\{\mathbf{z}_1, \ldots, \mathbf{z}_{i-1}, \mathbf{z}_{i+1}, \ldots, \mathbf{z}_K\}$. We consider three different objectives for learning the latent subspaces: 1) minimizing a supervised loss L (e.g., mean squared error or cross-entropy), $\sum_{i=1}^{K} L(\hat{s}_i, s_i)$, denoted *"Base"*; 2) minimizing the *unconditional mutual information between subspaces* in addition to the supervised loss, $\sum_i L(\hat{s}_i, s_i) + I(\mathbf{z}_1, \ldots, \mathbf{z}_K)$, denoted *"Base+MI"*; and 3) minimizing the *conditional mutual information between subspaces conditioned on observed attributes*, in addition to the supervised loss, $\sum_i L(\hat{s}_i, s_i) + I(\mathbf{z}_i; \mathbf{z}_{-i} \mid s_i)$ denoted *"Base+CMI"*. We wish to learn a model that is robust to correlation shifts, e.g., if we train on data where $\text{corr}(s_i, s_j) > 0$, then we desire that the resulting model will perform similarly on uncorrelated data, $\text{corr}(s_i, s_j) = 0$, or anticorrelated data, $\text{corr}(s_i, s_j) < 0$.

In this section, we motivate the use of CMI minimization for learning robust disentangled representations. We use a linear regression task that can be solved analytically, and for which all quantities of interest, including MI and CMI, can be computed in closed form. This allows us to compare the solutions obtained via the vanilla mean-squared error objective (*Base*) to the solutions obtained by minimizing the MSE *under the constraint* that the MI or CMI between latent subspaces is minimized. This yields insight into the behavior of the objectives in the idealized case where the constraints they prescribe ($I(z_1; z_2) = 0$ for MI or $I(z_1; z_2 \mid s_1) = I(z_1; z_2 \mid s_2) = 0$ for CMI) are exactly satisfied.

First, we show that the supervised loss alone does not yield robust disentangled representations. Then, we show that additionally minimizing the unconditional MI forces the model to learn an *even worse solution*. Finally, we show that minimizing the conditional MI yields appropriately disentangled representations that are robust to correlation shift.

3.1 Full Supervision Does Not Yield Disentanglement

Here, we introduce a linear regression problem with correlated attributes. First, we analyze the solution obtained by optimizing only the *Base* objective, which in this case is the mean squared error. Consider a linear generative model with correlated Gaussian source variables $\mathbf{s}$, given by:

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n} \quad , \quad \mathbf{s} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_s}) \quad , \quad \mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_n})$$

where $\mathbf{A}$ is the ground-truth mixing matrix and $\mathbf{C_s}$ and $\mathbf{C_n}$ are the covariance matrices for the source and noise variables, respectively. We assume that $\mathbf{x}$ is observed and wish to disentangle the underlying source variables $\mathbf{s}$; this corresponds to finding the mapping $\mathbf{A}^{-1}$ that inverts the data generating process. When we have access to the source variables, a natural approach is to minimize a supervised loss to ensure that each subspace contains information about its attribute. The optimal linear regression solution, both in the least squares sense and with respect to maximum likelihood, is given by the posterior mean:

$$\hat{\mathbf{s}}(\mathbf{x}) = \mathbb{E}\left[\mathbf{s} \mid \mathbf{x}\right] = \mathbf{C_{sx}}\mathbf{C_x^{-1}}\mathbf{x} \tag{1}$$

where $\mathbf{C_{sx}}$ and $\mathbf{C_x}$ are the following covariance matrices:

$$\mathbf{C_{sx}} = \mathbb{E}\left[\mathbf{s}(\mathbf{A}\mathbf{s} + \mathbf{n})^\top\right] = \mathbf{C_s}\mathbf{A}^\top \tag{2}$$

$$\mathbf{C_x} = \mathbf{A}\mathbf{C_s}\mathbf{A}^\top + \mathbf{C_n} \tag{3}$$

The least-squares optimal mapping $\mathbf{C_{sx}}\mathbf{C_x^{-1}}$ in Eq. 1 is not equal to the inverse $\mathbf{A}^{-1}$ of the generative model, as it is biased by the correlation structure $\mathbf{C_s}$ and $\mathbf{C_n}$ towards directions of maximal signal-to-noise ratio. Thus, regression is sensitive to noise, and this can lead to failures when evaluating the model on correlation-shifted data. For this Gaussian problem, we can compute the expected mean squared error (and therefore the expected variance explained) analytically:

$$\mathbb{E}\left[(\mathbf{s} - \hat{\mathbf{s}}(\mathbf{x}))^2\right] = \text{Var}\left(\mathbf{s}\right) = \text{Tr}(\mathbf{C_s}) \tag{4}$$

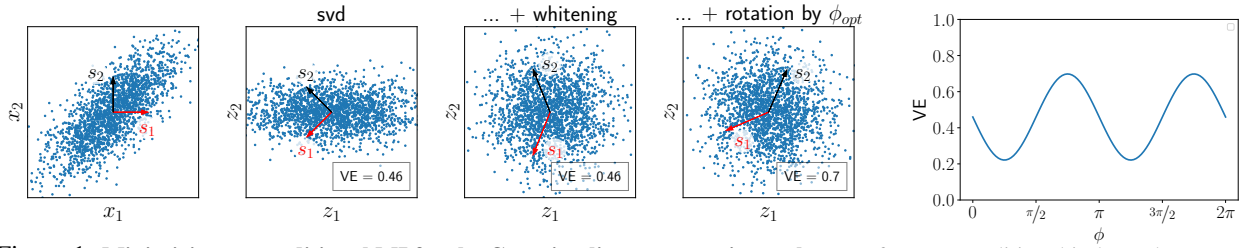


Figure 1: **Minimizing unconditional MI for the Gaussian linear regression task.** To enforce unconditional independence, we choose $\mathbf{W}$ such that $\mathrm{Cov}(\mathbf{z})$ is diagonal. In our case this is easy: the principal components of $\mathbf{x}$ are $x_1 + x_2$ and $x_1 - x_2$. The optimal regression loss with minimal MI is then given by whitening and rotating the result by angle ϕ_{opt} which leads to maximal variance explained ($\phi_{\text{opt}} = -\pi/4$ for positive correlations and $\mathbf{A} = \mathbf{I}$).

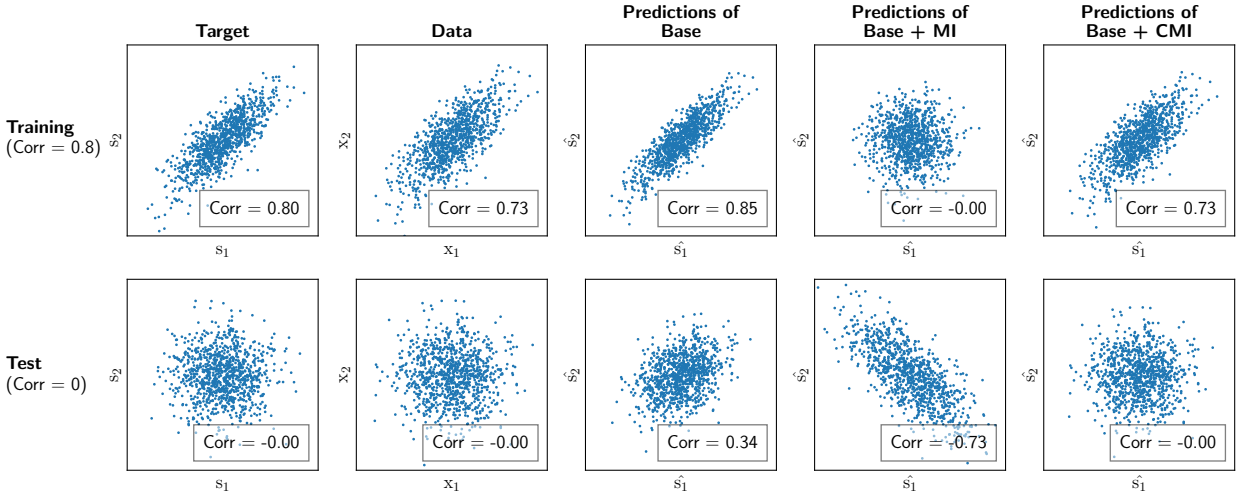


Figure 2: **Visualisation of targets $\mathbf{s}$, input data $\mathbf{x}$ and the predictions $\hat{\mathbf{s}}$ made by models using each of the different objectives {*Base*, *Base+MI*, *Base+CMI* }.** For *Base*, the predictions are more correlated than the data, revealing that the correlation in the training data is used to compensate for the noise. *Base+MI* leads to uncorrelated predictions. This cannot be the correct solution, as the targets are correlated. Only for *Base+CMI* does the correlation between the predictions and data match for both training and test data.

In Table 1, we see that in the two-dimensional case where $\mathbf{s} = (s_1, s_2)$ for $\mathbf{A} = \mathbf{I}$, $\mathbf{C_n} = 0.01 \cdot \mathbf{I}$ and the train-time correlation is $\mathrm{corr}(s_1, s_2) = 0.8$, $\hat{\mathbf{s}}$ explains 91.9% of the variance in $\mathbf{s}$ (column *"Base"*). However, when the correlation between s_1 and s_2 shifts at test time, such that $\mathrm{corr}(s_1, s_2) = 0$, then performance drops to 87.6%. This drop occurs because the estimator $\hat{\mathbf{s}}$ tries to make use of the assumed correlation between s_1 and s_2 to counteract the information lost due to noise, but this correlation is no longer present in the test data (see also Figure 2). The gap in performance between correlated and uncorrelated data indicates that s_1 and s_2 have not been correctly disentangled.

3.2 Unconditional Disentanglement Fails Under Correlation Shift

In the 2D linear case, we have:

$$\mathbf{z} = (z_1, z_2) = \mathbf{W}\mathbf{x}, \qquad \widehat{s}_1 = R_1 z_1, \qquad \widehat{s}_2 = R_2 z_2 \tag{5}$$

where the matrix $\mathbf{W}$ encodes the observation into the latent space. The linear regression example in Sec. 3.1 corresponds to $\mathbf{W} = \mathbf{C_{sx}}\mathbf{C_x^{-1}}$ and $R_k = 1$. In standard supervised objectives, there is no constraint preventing a subspace z_k from containing information about other source variables than s_k. A common approach to enforce independence is to minimize the MI between the latent subspaces z_1 and z_2 (Chen et al., 2018; Peng et al., 2019). In the Gaussian case, random variables are independent if and only if they are *uncorrelated*. The optimal linear regression weights $\mathbf{W}$ that yield $I(z_1; z_2) = 0$ (e.g., such that $\mathrm{Cov}(\mathbf{z})$ is diagonal) can be computed by whitening $\mathbf{x}$ and rotating the result by an angle ϕ_{opt} which leads to maximal variance explained. For our example in Table 1, where we have positive correlation and $\mathbf{A} = \mathbf{I}$, the optimal rotation is $\phi_{\text{opt}} = -\pi/4$ (see Figure 1). However, the resulting model no longer performs well on in-distribution data (Table 1, column *"Base+MI"*). There is correlation between the source variables s_1 and s_2 and therefore $I(s_1; s_2) > 0$. By enforcing independence, at least one of the subspaces cannot contain all relevant

information about its attribute and thus will have poor predictive performance. We make this precise in the following proposition.

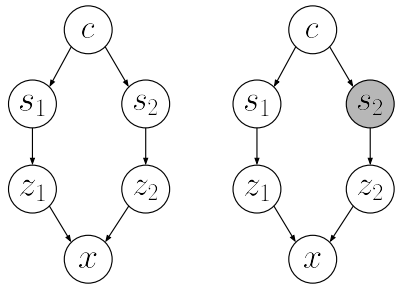
Proposition 3.1. *If $I(s_1; s_2) > 0$, then enforcing $I(z_1; z_2) = 0$ leads to $I(z_k; s_k) < H(s_k)$ for at least one k.*

Proof. The proof is provided in Appendix D. □

3.3 Conditional Disentanglement is Robust to Correlation Shift

We have seen that enforcing unconditional independence between the latent spaces does not solve the disentanglement problem. However, considering the graphical model in Figure 3, $\mathbf{z}_1$ and $\mathbf{z}_2$ are independent *conditioned on either of* s_1 *or* s_2: assuming a common cause for the correlation between s_1 and s_2, there is a connection in the graphical model between $\mathbf{z}_1$ and $\mathbf{z}_2$ introducing a statistical dependence. Observing either s_1 or s_2 disconnects $\mathbf{z}_1$ and $\mathbf{z}_2$. Here, we show that enforcing independence *conditioned on each of the source variables* is also sufficient to yield a robust disentangled representation. For our 2D example, enforcing conditional independence corresponds to:

$$\mathrm{I}(\mathbf{z}_1; \mathbf{z}_2 \mid s_1) = 0 \qquad \text{and} \qquad \mathrm{I}(\mathbf{z}_1; \mathbf{z}_2 \mid s_2) = 0 \tag{6}$$



$I(z_1; z_2) > 0$ $\quad$ $I(z_1; z_2 | s_2) = 0$

Figure 3: The graphical model for two sources s_1, s_2 and corresponding latent subspaces z_1, z_2. We assume the source variables have a common cause c. In (a), when none of the sources are observed, there is a path from z_1 to z_2, so we have $I(z_1; z_2) > 0$; in (b) we observe s_2, which breaks the path, and thus $I(z_1; z_2 \mid s_2) = 0$.

Intuitively, if s_1 and s_2 are correlated, then $I(s_1; s_2) > 0$ and knowing s_1 gives us information about s_2. If we can predict s_1 from $\mathbf{z}_1$, and s_1 tells us about s_2, then it must be the case that $\mathbf{z}_1$ contains information about s_2.

We wish to ensure that $\mathbf{z}_1$ and $\mathbf{z}_2$ share *as little information as possible* (given the ground-truth correlation), to improve robustness to shifts. Since $\mathbf{z}_1$ necessarily contains some information about s_2, we enforce that it does not contain *any more information about* $\mathbf{z}_2$ *than necessary* via $I(\mathbf{z}_1; \mathbf{z}_2 | s_2)$, which states that if we know s_2, then knowing $\mathbf{z}_1$ does not give us more information about $\mathbf{z}_2$.

This does not penalize $\mathbf{z}_1$ for containing information about s_2 due to correctly predicting the correlated variable s_1 (and vice versa). In contrast to MI, this removes only the shared information which is not robust under correlation shift, but keeps the shared information which is necessary to account for the correlation between the source variables. The optimal solution under the conditional independence constraint (Eq. 6) is achieved by the mapping $\mathbf{W} = \mathbf{A}^{-1}$, successfully recovering the underlying generative model. This demonstrates the usefulness of minimizing CMI for generalization under correlation shifts in the case of linear regression with Gaussian variables and motivates us to investigate CMI minimization for larger-scale tasks.

4 Method: Minimizing CMI

For simple cases such as linear regression, we can compute and minimize the MI and CMI analytically; however, for most tasks, there is no closed form for the mutual information. In this section, we describe an approach to minimize the CMI for general classification tasks. Suppose we have a dataset $\mathcal{D} = \{(\mathbf{x}^{(i)}, \mathbf{s}^{(i)})\}_{i=1}^{N}$ where $\mathbf{x}^{(i)}$ is an example and $\mathbf{s}^{(i)}$ is a vector of attribute labels — $\mathbf{s}_k^{(i)}$ is the label for the k^{th} attribute of the i^{th} example. We consider discrete attributes, $\mathbf{s}_k^{(i)} \in \mathbb{N}$. Let $f_{\boldsymbol{\theta}} : \mathbf{x} \mapsto \mathbf{z}$ denote an encoder parameterized by $\boldsymbol{\theta}$ that maps examples $\mathbf{x} \in \mathbb{R}^m$ to latent representations $\mathbf{z} \in \mathbb{R}^n$. We aim to learn one latent subspace per attribute, such that each subspace is independent from all other subspaces conditioned on the attribute it encodes.

We have $I(\mathbf{x}; \mathbf{y} \mid \mathbf{z}) = 0$ if $p(\mathbf{x}, \mathbf{y} \mid \mathbf{z}) = p(\mathbf{x} \mid \mathbf{z})p(\mathbf{y} \mid \mathbf{z})$. Our method enforces the latter condition using an adversarial discriminator. To obtain samples from $p(\mathbf{z}_1, \ldots, \mathbf{z}_K \mid \mathbf{s}_k)$ and $p(\mathbf{z}_k \mid \mathbf{s}_k)p(\mathbf{z}_{-k} \mid \mathbf{s}_k)$, we loop over values of $\mathbf{s}_k$, and for each condition $\{\mathbf{s}_k = 0, \mathbf{s}_k = 1, \ldots\}$, we select examples from the minibatch that satisfy the condition, giving us samples from $p(\mathbf{z}_1, \ldots, \mathbf{z}_K \mid \mathbf{s}_k)$; then we shuffle the latent subspaces $\mathbf{z}_j, \forall j \neq k$ jointly batchwise (e.g., combining $\mathbf{z}_k$ from one example with $\mathbf{z}_{-k}$ from another) to obtain samples from $p(\mathbf{z}_k \mid \mathbf{s}_k)p(\mathbf{z}_{-k} \mid \mathbf{s}_k)$. To enforce $p(\mathbf{z}_1, \ldots, \mathbf{z}_K \mid \mathbf{s}_k) = p(\mathbf{z}_k \mid \mathbf{s}_k)p(\mathbf{z}_{-k} \mid \mathbf{s}_k)$, we train the encoder f adversarially against a discriminator trained to distinguish between these two distributions. The discriminator takes as input a representation and predicts whether it is "real" (e.g., drawn from the joint distribution) or "fake" (e.g., drawn from the product of marginals). One discriminator is trained for each attribute $\mathbf{s}_k$, which receives samples from the two distributions and the attribute value it is conditioned on. In practice, we use a conditional discriminator, effectively sharing parameters between the discriminators for each of the attributes. This process is illustrated in Figure 4. Algorithm 1 describes the encoder training loop; Algorithm 5

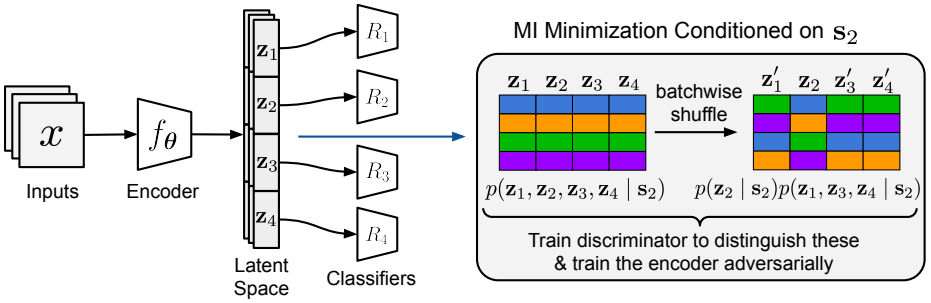


Figure 4: **Adversarial minimization of conditional mutual information via latent-space shuffling.** We minimize the CMI between latent subspaces, $I(\mathbf{z}_1; \cdots; \mathbf{z}_K \mid \mathbf{s}_k)$. Here, we illustrate the algorithm for four attributes with corresponding latent spaces $\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3, \mathbf{z}_4\}$, where we condition on attribute $\mathbf{s}_2$. See Sec. 4 for a description of the method.

Algorithm 1 Adversarial Learning of Conditionally Disentangled Subspaces — Training the Encoder

Input: $\{\phi_1, \ldots, \phi_K\}$, initial parameters for K linear classifiers $R_1, \ldots, R_K$
Input: $\boldsymbol{\theta}$, initial parameters for the encoder f
Input: α, β learning rates for training the encoder and linear classifiers
while true **do**
  $(\mathbf{x}, \{\mathbf{s}_k\}_{k=1}^K) \sim \mathcal{D}_{\text{Train}}$ ▷ Sample a minibatch of data with attribute labels
  $\mathbf{z} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{x})$ ▷ Forward pass through the encoder
  $\{\mathbf{z}_k\}_{k=1}^K \leftarrow \text{SplitSubspaces}(\mathbf{z}, K)$ ▷ Partition the latent space into K subspaces
  $L \leftarrow \sum_{k=1}^K L_{\text{cls}}(R_k(\mathbf{z}_k; \boldsymbol{\phi}_k), \mathbf{s}_k)$ ▷ Cross-entropy for each attribute
  for $k \in \{1, \ldots, K\}$ **do** ▷ For each attribute/subspace
    $\mathbf{z}' \sim p(\mathbf{z}_1, \ldots \mathbf{z}_K \mid \mathbf{s}_k)$ ▷ Samples from the joint distribution
    $\mathbf{z}'' \sim p(\mathbf{z}_k \mid \mathbf{s}_k)p(\mathbf{z}_{-k} \mid \mathbf{s}_k)$ ▷ Samples w/ batchwise-shuffled subspaces
    $L \leftarrow L + \log\left(1 - D_{\boldsymbol{\omega}}(\mathbf{z}'')\right) + \log\left(D_{\boldsymbol{\omega}}(\mathbf{z}')\right)$ ▷ Add adversarial loss
  end for
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} L$ ▷ Update encoder parameters
  $\boldsymbol{\phi}_k \leftarrow \boldsymbol{\phi}_k - \beta \nabla_{\boldsymbol{\phi}_k} L \quad , \quad \forall k \in \{1, \ldots, K\}$ ▷ Update classifier parameters
end while

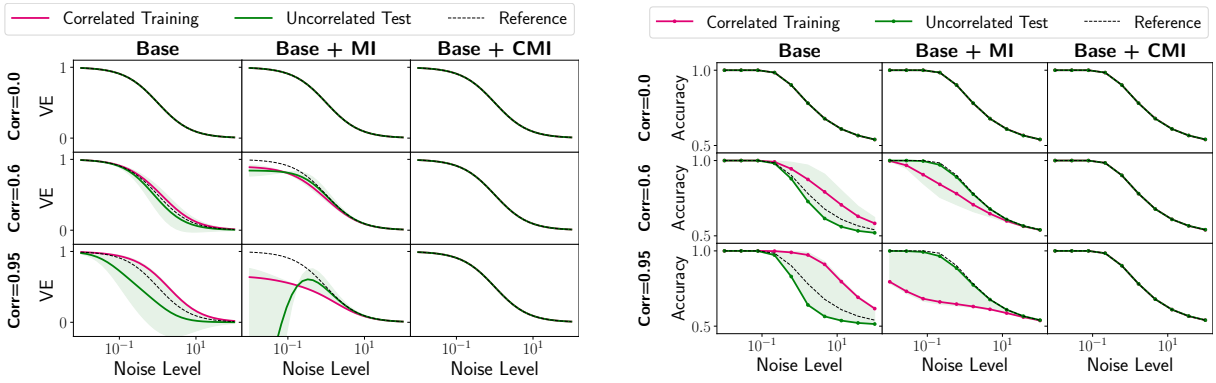
in Appendix C describes the corresponding discriminator training loop. We formally describe the algorithms for the baselines (*Base* and *Base + MI*) in Appendix C.

This approach is architecture-agnostic, and can be used to factorize the latent space of any classifier or generative model (e.g., VAEs (Joy et al., 2020) or flow-based models (Kingma & Dhariwal, 2018)). However, some models (such as VAEs) may have objectives that interfere with the goal of obtaining conditionally independent subspaces; for example, the ELBO encourages independence between all latent dimensions. In our experiments, we used linear and MLP encoders rather than VAEs to avoid this conflicting objective.

Because the latent space is typically low-dimensional, we have a choice of different distribution alignment techniques, including maximum mean discrepancy (MMD) (Gretton et al., 2006) and adversarial approaches (Goodfellow et al., 2014). Different GAN formulations can be interpreted as minimizing different divergences: the vanilla GAN (Goodfellow et al., 2014) minimizes the Jensen-Shannon divergence; WGAN (Arjovsky et al., 2017) minimizes the Wasserstein distance, which has been used to define an analogue of mutual information called the *Wasserstein dependency measure* (Ozair et al., 2019); f-GAN (Nowozin et al., 2016) minimizes an arbitrary f-divergence, etc. Each of these divergence measures will be 0 if and only if the subspaces are independent, however their training dynamics may differ. In practice, we found the vanilla GAN formulation to work well across our experiments.

5 Experiments

Our experiments aim to answer the following questions: 1) What is the effect of the train-time correlation strength and noise level on the solutions found by training with each objective, *Base*, *Base+MI*, and *Base+CMI*? 2) Can we successfully learn conditionally disentangled representations for classification tasks using Algorithm 1? and 3) Does CMI minimization lead to improved correlation-shift robustness on natural image datasets including MNIST and CelebA?



(a) Toy linear regression.

(b) Toy classification with ten attributes.

Figure 5: **Synthetic linear regression (left) and linear classification (right) tasks.** We measure the performance (variance explained for regression and accuracy for classification) on the correlated training data (magenta) and on test data with a range of correlation shifts (green, solid line is the uncorrelated test data). The performance of the *Base* model in the uncorrelated setting serves as a reference in each plot (dashed black line) and facilitates the comparison of the performance of the different objectives (columns). In both tasks, we find that, *Base+CMI* leads to robustness to correlation shift independent of the noise level (x-axis) and the strength of the correlation in the training data (rows), while the other approaches do not.

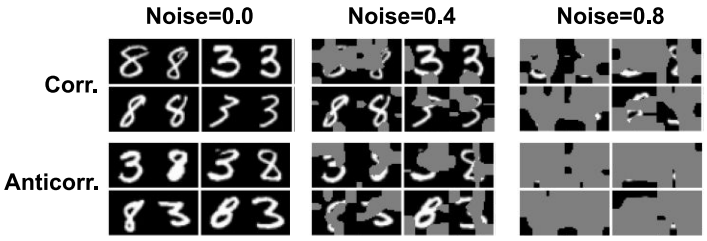
First, we present results on the analytically-solvable linear regression example, illustrating the effect of the correlation strength and noise level on the solution obtained by each objective. Then, we demonstrate that our findings also hold for a synthetic classification task with multiple attributes. Next, we employ the method described in Section 4 and investigate two realistic tasks, a multi-digit MNIST task with occlusions and correlated CelebA, and show that minimizing CMI can largely eliminate the gap in performance caused by test-time correlation shifts. Finally, we evaluate common disentanglement metrics and apply Algorithm 1 in weakly supervised settings. Experimental details and extended results are provided in Appendix B.

Linear Regression. Here, we revisit the linear regression problem from Section 3, to investigate the impact of the train-time correlation strength and noise level on the models learned with each of the objectives *Base*, *Base+MI*, and *Base+CMI*. The results are shown in Figure 5a. We found that *Base+CMI* yields robustness to correlation shift across all correlation strengths and noise levels, while the baselines do not. The performance of *Base* drops most severely under correlation shift for strong train-time correlations and intermediate noise levels; in this regime, *Base+CMI* improves performance substantially.

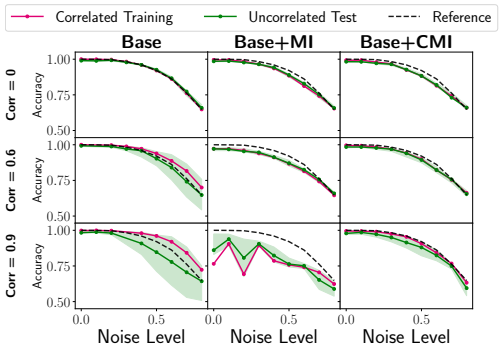
Toy Multi-Attribute Classification. Next, we investigated whether these findings hold for classification tasks with multiple attributes. Here, binary source attributes $s_k = \pm 1, \forall k \in \{1, \dots, K\}$ generate the observed data via $\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{n}$ (we set $\mathbf{A} = \mathbf{I}$ for simplicity) with normally distributed noise $\mathbf{n} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_n})$. We induced correlations between the attributes a_k, such that the number of datapoints differs for the different combinations of attribute values. In the multi-attribute setting, the correlation strength refers to the pairwise correlation between all attributes. Similarly to the regression task, we find that *Base+CMI* leads to robustness under correlation shift (see Figure 5b and Appendix B.1).

Multi-Digit Occluded MNIST. Next, we designed a larger-scale task to investigate whether these properties hold in a more complex setting. We created a dataset by concatenating two MNIST digits side-by-side, where the aim is to predict both the left- and right-hand labels. We generated occlusion masks using the procedure used by Chai et al. (2021); examples from our synthetic dataset under a range of noise settings are shown in Figure 6a. We used a subset of MNIST consisting of classes `3` and `8` (which are visually similar and can become ambiguous under occlusions). This mimics multiple-object classification in a way that allows us to control the correlation strength and noise level (via the amount of occlusion), allowing for systematic analysis. This task is a more complex analogue of the synthetic classification task from Figure 5b. We added explicit occlusion noise because the MNIST data itself is simple, and has too little "natural" noise to clearly observe the predicted effects (e.g., for low noise levels, the supervised loss already does well). While this task would also be possible for colored MNIST and dSprites, one advantage of our task is its symmetry, which allows us to exclude potential side-effects: here, the attributes have the same type (the digit identity), whereas the attributes in colored MNIST (digit identity and color) and dSprites (shape, size, position, etc.) are more diverse.

Similarly to the toy tasks, we train an encoder to map images onto a D-dimensional latent space, which is partitioned in two equal-sized subspaces corresponding to the two digits; we train a linear classifier on each subspace to predict the

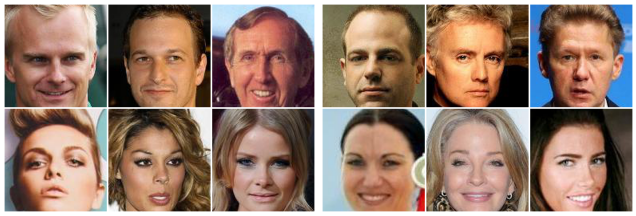


(a) Correlated training data and anti-correlated test data.



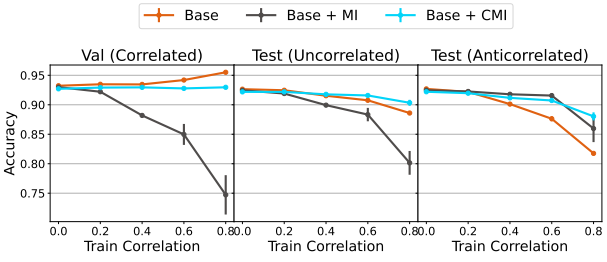
(b) Performance comparison.

Figure 6: **Multi-digit occluded MNIST.** (a) Examples of the correlated training data (where `3-3` and `8-8` pairs are frequent) and anticorrelated test data (where `3-8` and `8-3` pairs are frequent), under a range of occlusion strengths. (b) Accuracies under correlation shifts for different noise levels, achieved by training with each of the objective functions *Base*, *Base+MI*, and *Base+CMI*. *Base+CMI* achieves consistent performance across correlation shifts. Similarly to Figure 5, here we show the reference performance of the model trained on uncorrelated data (solid black line), the performance on correlated training data (magenta) and on a range of test-time correlations in $[0, 1]$ (shaded green region, where solid green denotes the uncorrelated test performance).



(a) Correlated train data.

(b) Anti-correlated test data.



(c) Performance comparison.

Figure 7: **Correlated CelebA.** (a) Training examples with correlation 0.8 between attributes `Male` and `Smiling`, such that the majority of men are smiling while the majority of women are not. (b) Anti-correlated test examples, where the majority of women are smiling. (c) Accuracies of each method under a range of correlation strengths, for validation data with the same correlation as the training data, uncorrelated test data, and anticorrelated test data.

respective class labels. We consider different correlation strengths between the left and right digits in the training set (where strong correlation means that the digits often match, e.g., `3-3` or `8-8` are more common than `3-8` or `8-3`). We evaluate each model on test data with correlation strengths ranging from $[-1, 1]$. The results are shown in Figure 6b. We found that the conclusions from the toy experiments hold in this setting: supervised learning with only the cross-entropy loss, as well as with unconditional MI minimization, fail under test-time correlation shift, while minimizing CMI is more robust. Experimental details and extended results are provided in Appendix B.2.

Correlated CelebA. Finally, we consider a realistic setting using the CelebA faces dataset (Liu et al., 2015). In contrast to the multi-digit MNIST task, here we do not add any artificial observation noise (as CelebA is a more complex dataset that naturally has noise in observations and/or labels). We selected two attributes that we know *a priori* are not causally related, `Male` and `Smiling`, and we created subsampled datasets with a range of training correlations $\{0, 0.2, 0.4, 0.6, 0.8\}$. We evaluated our models on both *anti-correlated* and *uncorrelated* test sets (Figures 7a and 7b). Figure 7c compares the performance of the baseline classifier, unconditional MI model, and conditionally disentangled model under a range of correlation strengths. We found that minimizing CMI has a larger effect for medium-to-high correlation; however, CMI minimization does not hurt performance at low correlation strengths. Note that while the unconditional model appears to have good performance on the anti-correlated test set, its performance is poor on the validation set (that has the same correlation structure as the training set), so this model does not perform well on in-distribution-data. In contrast, the *Base+CMI* model performs well on both in-distribution data and shifted test distributions. Also note that the problem of disentangling correlated attributes does not occur only under correlation shift, but is already present in the source domain where certain attribute combinations will reliably be treated incorrectly. For example, *Base* fails to recognize the rare non-smiling male faces in 49% of the cases, while *Base+CMI* fails only in 25% of the cases. Additional details are in Appendix B.3.

Disentanglement Metrics. Locatello et al. (2020a) showed that common disentanglement metrics are not suitable for the correlated setting. For this reason, we focused on comparing performance under correla-

tion shift, which we consider more suitable for correlated data: if a model cannot predict a factor of variation well for certain values of another factor, then the model did not successfully disentangle those factors. However, one can still make use of the disentanglement metrics by evaluating them on *uncorrelated data*, using models trained on correlated data. We performed this analysis for the toy classification and CelebA tasks, and found that *Base+CMI* leads to improved disentanglement scores across a wide range of metrics, compared to *Base* and *Base+MI* (Appendix B.4).

Extension to the Weakly Supervised Setting. Algorithm 1 can be applied directly to weakly supervised settings; it is not necessary for each datapoint to have labels for all attributes. We find that when reducing the number of labels, *Base+CMI* outperforms the other objectives under correlation shift (see Figure 8 and Appendix B.5).

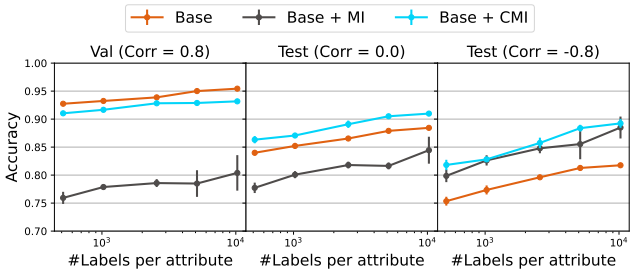


Figure 8: **Weakly-supervised CelebA.** The x-axis shows the number of labels per attribute used during training; the rightmost datapoint corresponds to full supervision. *Base+CMI* outperforms the other objectives under correlation shift.

6 Limitations & Future Work

Our study mainly concerns the setting where the underlying factors of variation are known. Practical applications of this setting can occur with respect to fairness, where one may wish to train a model such that correlations that exist in the training data are not relied upon for prediction. Nonetheless, full supervision is a strong assumption and an exciting goal for future work would be to look into relaxing this assumption. Our experiment with the weakly supervised version of the CelebA experiment is a first step in this direction.

We have shown that minimizing CMI yields predictions that disregard correlations between attributes in the training data, which is helpful when correlations shift between the training and test data. This approach relies on knowing a priori which correlations should not be used. This is the case, for example, for fairness applications where a person's race or gender should not affect the results. A direction for future work would be to automatically determine which correlations are more or less likely to shift in held-out data and to add this step before applying our approach of avoiding the unwanted correlations. One may incorporate ideas from IRM (Arjovsky et al., 2019), which leverages multiple environments at training time to discover which correlations tend to shift and which are stable—e.g., to distinguish between causal and spurious correlations, the latter of which we wish to avoid relying on. A fruitful direction for future work would be to combine IRM-style discovery of spurious correlations with our approach, which can be used to control for these correlations when learning disentangled representations. In a related vein, there has been recent work which aims to discover environments when none are given explicitly (Creager et al., 2021), which may be useful in combination with our work.

While CMI is defined for both continuous and discrete attributes, our method of shuffling the latent subspaces is only applicable to discrete attributes. Discrete attributes are prevalent in many settings: in domain adaptation, the class and domain are discrete; in multi-object classification, the class of each object is a discrete attribute; the foreground and background of natural images are discrete, etc. Nevertheless, finding methods to minimize CMI for continuous attributes is an interesting direction for future work. Another caveat of our method for minimizing the CMI via latent subspace shuffling is the increased computational cost relative to minimizing the unconditional MI: the cost for CMI scales linearly with the number of attributes and attribute values, while the cost for MI is constant.

7 Conclusion

Correlations are prevalent in real-world data, yet pose a substantial challenge for disentangled representation learning. Standard approaches learn to rely on these correlations, especially when data are noisy, as the correlations provide an easy-to-learn signal with predictive power. When the attributes are not causally related, this leads to poor performance under test-time correlation shift. Although for small correlations the effects may not be large, relying on these correlations and thereby systematically treating a subset of the data incorrectly, can be catastrophic for fairness. We first showed that supervised learning and *unconditional* mutual information minimization fail to learn representations robust to such shifts. We then argued that the correct notion of disentanglement in such cases is *conditional disentanglement*, and we proposed a simple approach to minimize the conditional mutual information between latent subspaces. We showed that conditionally disentangled representations improve robustness to correlation shift in analytically solvable linear tasks, as well as on natural images. Overall, we established CMI minimization as a more appropriate alternative to MI minimization, which sets the stage for the development of more powerful objective functions for disentanglement.

ACKNOWLEDGEMENTS

We thank Jörn-Henrik Jacobsen for his valuable contributions in the early stage of this work. We thank Steffen Schneider, Dylan Paiton, Lukas Schott, Elliot Creager, and Frederik Träuble for helpful discussions. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Christina Funke. Paul Vicol was supported by a JP Morgan AI Fellowship.

We acknowledge support from the German Federal Ministry of Education and Research (BMBF) through the Competence Center for Machine Learning (FKZ 01IS18039A) and the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002), the German Excellence Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307), and the Deutsche Forschungsgemeinschaft (DFG; Projektnummer 276693517 – SFB 1233). Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/partners.

REFERENCES

Tameem Adel, Zoubin Ghahramani, and Adrian Weller. Discovering interpretable representations for both deep generative and discriminative models. In *International Conference on Machine Learning (ICML)*, pp. 50–59, 2018.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*, 2016.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning (ICML)*, pp. 214–223, 2017.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, pp. 456–473, 2018.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. MINE: Mutual information neural estimation. In *International Conference on Machine Learning (ICML)*, 2018.

Anthony J Bell and Terrence J Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton University Press, 2009.

Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with Imagenet? *arXiv preprint arXiv:2006.07159*, 2020.

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

Chris Burgess and Hyunjik Kim. 3D Shapes Dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

J-F Cardoso. Multidimensional independent component analysis. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pp. 1941–1944, 1998.

Lucy Chai, Jonas Wulff, and Phillip Isola. Using latent space regression to analyze and leverage compositionality in GANs. *arXiv preprint arXiv:2103.10426*, 2021.

Agisilaos Chartsias, Thomas Joyce, Giorgos Papanastasiou, Scott Semple, Michelle Williams, David Newby, Rohan Dharmakumar, and Sotirios A Tsaftaris. Factorised spatial representation learning: Application in semi-supervised myocardial segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 490–498, 2018.

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

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. Club: A contrastive log-ratio upper bound of mutual information. In *International Conference on Machine Learning (ICML)*, pp. 1779–1788, 2020.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using mutual information. In *International Symposium on Information Theory (ISIT)*, pp. 2521–2526. IEEE, 2020.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3606–3613, 2014.

Pierre Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589*, 2019.

Elliot Creager, Jörn-Henrik Jacobsen, and Richard Zemel. Environment inference for invariant learning. In *International Conference on Machine Learning*, pp. 2189–2200. PMLR, 2021.

Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wüthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. *arXiv preprint arXiv:2010.14407*, 2020.

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain Markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 36(2):183–212, 1983.

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

Zunlei Feng, Xinchao Wang, Chenglong Ke, An-Xiang Zeng, Dacheng Tao, and Mingli Song. Dual swap disentangling. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5894–5904, 2018.

Behnam Gholami, Pritish Sahu, Ognjen Rudovic, Konstantinos Bousmalis, and Vladimir Pavlovic. Unsupervised multi-target domain adaptation: An information theoretic approach. *IEEE Transactions on Image Processing*, 29: 3993–4002, 2020.

Muhammad Waleed Gondal, Manuel Wüthrich, Đorđe Miladinović, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: A new disentanglement dataset. *arXiv preprint arXiv:1906.03292*, 2019.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample problem. *Advances in Neural Information Processing Systems (NeurIPS)*, 19:513–520, 2006.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016.

Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. AttGAN: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

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. In *International Conference on Learning Representations (ICLR)*, 2017a.

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. *arXiv preprint arXiv:1707.08475*, 2017b.

Aapo Hyvärinen and Patrik Hoyer. Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. DIVA: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, pp. 322–348, 2020.

Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. *arXiv preprint arXiv:1811.00401*, 2018.

Tom Joy, Sebastian Schmon, Philip Torr, N Siddharth, and Tom Rainforth. Capturing label characteristics in VAEs. In *International Conference on Learning Representations (ICLR)*, 2020.

Christian Jutten and Jeanny Herault. Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, 1991.

Christian Jutten and Juha Karhunen. Advances in nonlinear blind source separation. In *International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pp. 245–256, 2003.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication*, pp. 1–6. IEEE, 2009.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *11th International Conference on Data Mining Workshops*, pp. 643–650. IEEE, 2011.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, pp. 2649–2658, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *arXiv preprint arXiv:1807.03039*, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

David Klindt, Lukas Schott, Yash Sharma, Ivan Ustyuzhaninov, Wieland Brendel, Matthias Bethge, and Dylan Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. *arXiv preprint arXiv:2007.10930*, 2020.

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

Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pp. 11–104, 2004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. PacGAN: The power of two samples in generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. *arXiv preprint arXiv:1809.01361*, 2018.

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

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14611–14624, 2019a.

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 *International Conference on Machine Learning (ICML)*, pp. 4114–4124, 2019b.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. A sober look at the unsupervised learning of disentangled representations and their evaluation. *Journal of Machine Learning Research*, 21:1–62, 2020a.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning (ICML)*, 2020b.

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

Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

Sina Molavipour, Germán Bassi, and Mikael Skoglund. Conditional mutual information neural estimator. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5025–5029, 2020a.

Sina Molavipour, Germán Bassi, and Mikael Skoglund. On neural estimators for conditional mutual information using nearest neighbors sampling. *arXiv preprint arXiv:2006.07225*, 2020b.

Arnab Kumar Mondal, Arnab Bhattacharya, Sudipto Mukherjee, Sreeram Kannan, Himanshu Asnani, and Prathosh AP. C-MI-GAN: Estimation of conditional mutual information using MinMax formulation. *arXiv preprint arXiv:2005.08226*, 2020.

Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. CCMI: Classifier based conditional mutual information estimation. In *Uncertainty in Artificial Intelligence*, pp. 1083–1093, 2020.

Jozsef Nemeth. Adversarial disentanglement with grouped observations. In *34th AAAI Conference on Artificial Intelligence*, 2020.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. F-GAN: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. *arXiv preprint arXiv:1903.11780*, 2019.

Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Advances on Neural Information Processing Systems (NeurIPS)*, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. In *International Conference on Machine Learning (ICML)*, 2019.

Ken Perlin. Improving noise. In *29th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 681–682, 2002.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019.

Aloys Leo Prinz. Chocolate consumption and Nobel laureates. *Social Sciences & Humanities Open*, 2(1):100082, 2020. ISSN 2590-2911. doi: 10.1016/j.ssaho.2020.100082. URL https://www.sciencedirect.com/science/article/pii/S2590291120300711.

Scott E Reed, Yi Zhang, Yuting Zhang, and Honglak Lee. Deep visual analogy-making. *Advances in Neural Information Processing Systems (NeurIPS)*, 28:1252–1260, 2015.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. *arXiv preprint arXiv:1910.09772*, 2019.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning (ICML)*, pp. 6056–6065, 2019.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Anirudh Goyal, Francesco Locatello, Bernhard Schölkopf, and Stefan Bauer. Is independence all you need? On the generalization of representations learned from correlated data. *arXiv preprint arXiv:2006.07886*, 2020.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. From Imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning (ICML)*, pp. 9625–9635, 2020.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7167–7176, 2017.

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 (NeurIPS)*, pp. 14245–14258, 2019.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML)*, pp. 325–333, 2013.

Cheng Zhang, Kun Zhang, and Yingzhen Li. A causal view on robustness of neural networks. *arXiv preprint arXiv:2005.01095*, 2020.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019.

APPENDIX

This appendix is structured as follows:

- In Section A we provide an overview of the notation we use throughout the paper.
- In Section B we provide experimental details, as well as extended results.
- In Section C we provide the algorithms for the baseline methods, namely for classification-only training and unconditional mutual information minimization.
- In Section D we provide a proof of Proposition 3.1.

A NOTATION

Symbol	Meaning
$\mathbf{x}$	Observations
$\mathbf{s}$	Ground-truth latent factors
$\hat{\mathbf{s}}$	Predictions of factors
$\mathbf{z}$	Latent representation
$\mathbf{W}$	Linear regression weights
R_1, R_2	Linear readout from the latent space $\mathbf{z}$ to predictions $\hat{\mathbf{s}}$
$\mathbf{n}$	Isotropic Gaussian noise, $\mathbf{n} \sim \mathcal{N}(0, \mathbf{C_n})$ with $\mathbf{C_n} = \sigma^2 I$
$\mathbf{A}$	Square matrix used to generate observations for the linear task as $\mathbf{x} = \mathbf{As} + \mathbf{n}$
f	Encoder function
f_θ	Encoder function with parameters θ

Table 2: Summary of the notation used in this paper.

B EXPERIMENTAL DETAILS AND EXTENDED RESULTS

Method Details. Note that the dimensions m and n are arbitrary—in particular, n does not need to be smaller than m. In principle, each subspace can have different dimension (e.g., the linear readout layer for each attribute can have arbitrary dimensions $A \times S$ where A is the attribute dimensionality and S is the dimensionality of a particular subspace).

Compute Environment. Our experiments were implemented using PyTorch (Paszke et al., 2019), and were run on our internal clusters. The toy 2D experiments were run on a single NVIDIA RTX 2080 TI GPU, and took approximately 48 hours for all the results presented. The MNIST and CelebA experiments were run on NVIDIA Titan Xp GPUs. Each run of the multi-digit MNIST and CelebA tasks for a given method and correlation strength (and noise level in the MNIST case) took approximately 12 hours, and these were run in parallel.

B.1 TOY MULTI-ATTRIBUTE CLASSIFICATION

We performed this experiment with two, four and ten binary attributes. The results for varying numbers of attributes are shown in Figure 10. For two attributes we illustrated the data $\mathbf{x}$ for different correlation strength and noise levels (Figure 9). Here, increasing the correlation strength means that data points with $a_1 = a_2$ are increasingly more common relative to $a_1 \neq a_2$. The noise level on the other hand determines the overlap of the distributions and therefore the difficulty of the task.

Experimental Details. We used a PacGAN-style setup (Lin et al., 2018) for our toy experiments, where the discriminator takes as input a concatenation of 50 samples.

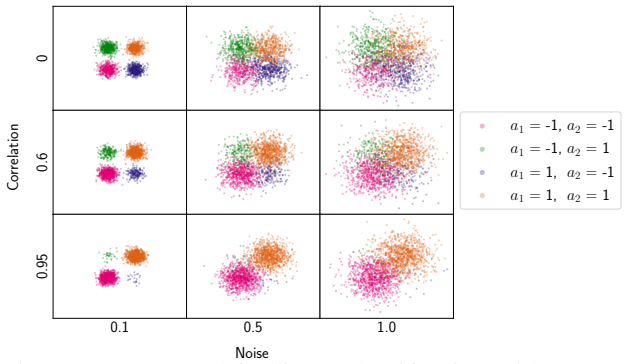
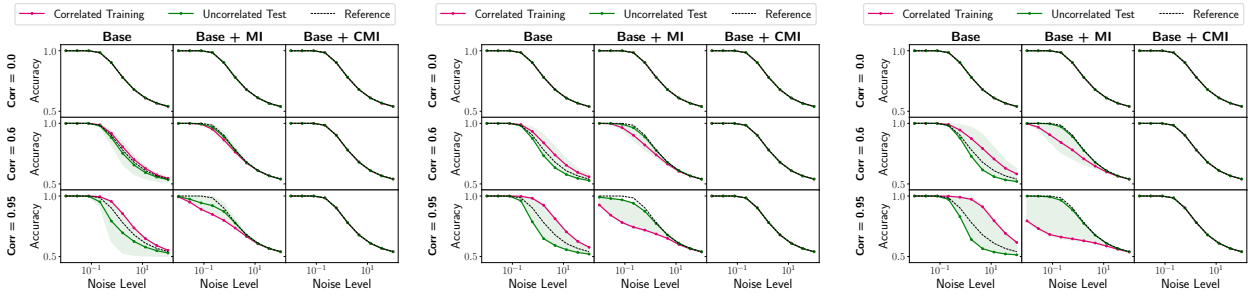


Figure 9: Data used for linear classification with two attributes (a_1 and a_2), visualized for a range of correlation strengths and noise levels.



(a) Two attributes. (b) Four attributes. (c) Ten attributes.

Figure 10: Toy classification with different numbers of attributes. Strong negative correlations could not be generated for multiple attributes; thus only positive test correlations were evaluated for (b) and (c).

- **Base**: We used Adam (Kingma & Ba, 2014) with a learning rate of 0.01.
- **Base + MI**: We used Adam to optimize the encoder, linear classifiers, and discriminators. After each step of optimizing the discriminator and encoder, we optimized the linear classifiers (R) for 10 steps. The disentanglement loss term was weighted by a factor of 100 relative to the classification loss. In preliminary tests, we found that the optimal learning rate depended on noise level, correlation strength, and number of attributes. The results in Figure 5b were obtained using one of the following learning rates for the discriminator $\{1e-4, 2e-4, 5e-4, 1e-3, 5e-3\}$. The learning rate of the generator and linear classifiers was chosen to be 10 times smaller than the discriminator learning rate.
- **Base + CMI**: For $\mathbf{A} = \mathbf{I}$, no optimization was necessary, as we already know the optimal solution to be $\mathbf{W} = \mathbf{A}^{-1} = \mathbf{I}$. We confirmed experimentally that the discriminator could not get above chance performance for this solution.

B.2 Multi-Object Occluded MNIST

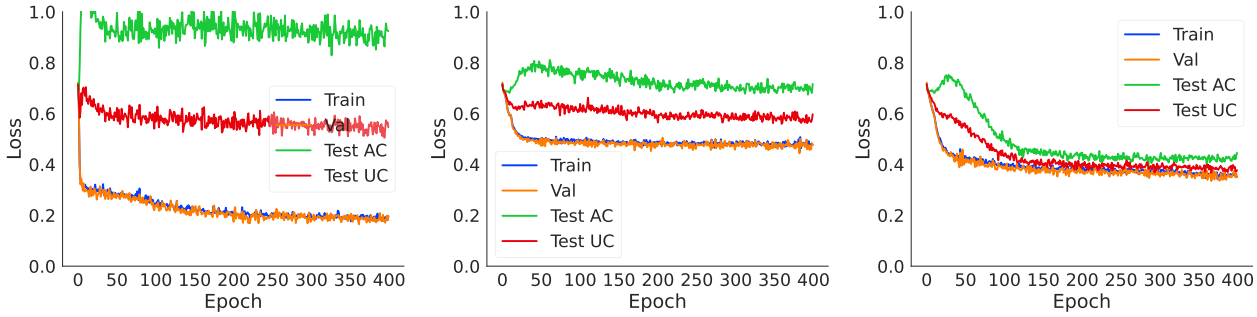
We used minibatch size 100, and latent dimension $D = 10$, yielding two subspaces each of dimension 5. As the encoder model, we used a three-layer MLP with 50 hidden units per layer and ReLU activations. We trained for 400 epochs, using Adam (Kingma & Ba, 2014) to optimize the encoder, linear classifiers, and discriminators, with separate learning rates for each component chosen via a grid search over $\{1e-5, 1e-4, 1e-3\}$.

Correlated Data Generation. We used the default MNIST training and test splits, and held out 10k of the original training examples to form a validation set, yielding 50k, 10k, and 10k examples in the training, validation, and test sets, respectively. Each digit is first rescaled to be 32×32 pixels. The correlated data was generated on-the-fly during training. Each example in a minibatch was created by: 1) sampling the left-right digit combination (e.g., { `3-3`, `3-8`, `8-3`, `8-8` }) from a joint distribution encoding the desired correlation; 2) choosing random instances of each of the selected classes (e.g., a random image of a `3` and a random image of an `8`); 3) applying occlusions separately to each image; and 4) concatenating the images, yielding a 32×64 example. This procedure was performed for each training and test minibatch, yielding a larger amount of data than would be possible with a fixed dataset generated *a priori*. To generate occlusions, we use the approach from (Chai et al., 2021), which produces contiguous masks similar to Perlin noise (Perlin, 2002). We used gray occlusions to remove a potential ambiguity that exists with black masks (which blend into the MNIST background): a masked `8` can become identical to a `3`, so one could not tell whether the image is a noisy `8` or a clean `3`.

B.3 CelebA

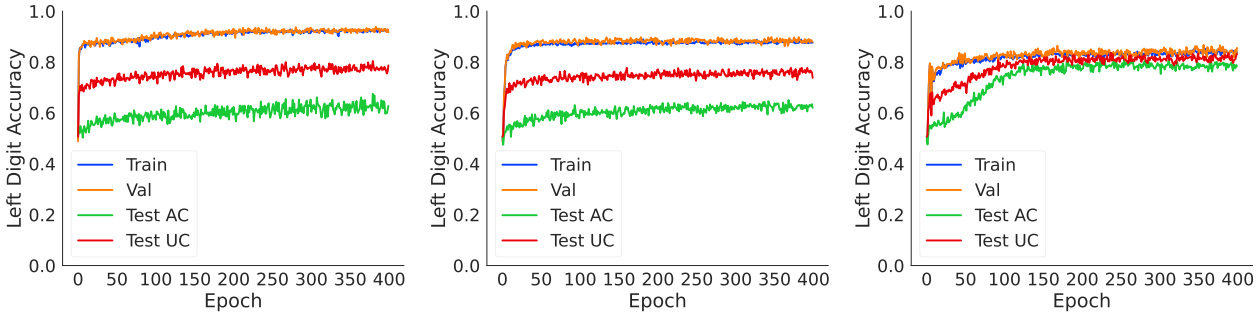
For all experiments, we used minibatch size 100, and latent dimension $D = 10$. As the encoder model, we used a three-layer MLP with 50 hidden units per layer and ReLU activations. Similarly to the MNIST setup, we trained for 200 epochs, using Adam to optimize the encoder, linear classifiers, and discriminators. For each method, we performed a grid search over learning rates $\{1e-5, 1e-4, 1e-3\}$ separately for each of the encoder, discriminator(s), and linear classification heads; we selected the best learning rates based on validation accuracy.

Correlated Data Generation. We first pre-processed all images by taking a 128×128 center crop, and then resizing to 64×64. Pixel values were normalized to the range $[0, 1]$. We used the original training, validation, and test splits



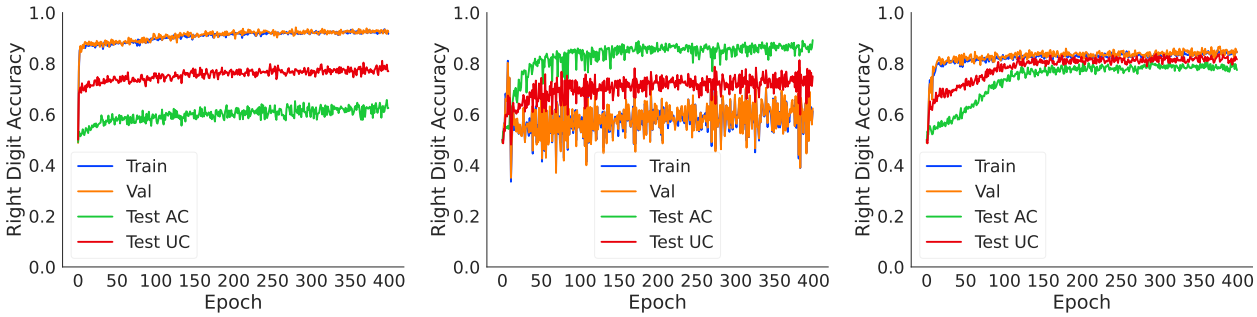
(a) Base (b) Base + MI (c) Base + CMI

Figure 11: Average cross-entropy loss for the left and right digit predictions, under the strongest correlation we consider, $c = 0.9$, at noise level 0.6 (where the noise is parameterized by a factor that has range $[0, 1]$).



(a) Base (b) Base + MI (c) Base + CMI

Figure 12: Accuracies for the left digit, under the strongest correlation we consider, $c = 0.9$, at noise level 0.6 (where the noise is parameterized by a factor that has range $[0, 1]$).



(a) Base (b) Base + MI (c) Base + CMI

Figure 13: Accuracies for the right digit, under the strongest correlation we consider, $c = 0.9$, at noise level 0.6 (where the noise is parameterized by a factor that has range $[0, 1]$).

provided with the CelebA dataset. In order to enforce arbitrary correlations between specific attributes, we subsampled the data such that we retained the maximum possible number of examples in each of the Train/Validation/Anticorrelated Test/Uncorrelated Test splits, while satisfying precisely the desired correlation. The validation set has the same correlation as the training set, and Figure 14 shows the number of examples in each of these sets for the strongest correlation we consider, $c = 0.8$. Figures 15, 16, and 17 show the cross-entropy loss and accuracies on each of the factors `Male` and `Smiling` (with training correlation 0.8) over the course of optimization, for each of the methods we compare (classification-only, unconditional disentanglement, and conditional disentanglement). We see that the conditional model substantially outperforms the baselines, with a much smaller gap between validation accuracy and both anti-correlated (AC) and uncorrelated (UC) test accuracies. Figures 18, 19 and 20 show confusion matrices for each method on the correlated validation set, anticorrelated test set, and uncorrelated test set, respectively. Finally,

Tables 3 and 4 show the prediction error of the models trained with the different objectives for both the combinations that were common and rare during training. These results shows that some attribute combinations (such as the rare non-smiling male faces) are reliably treated incorrectly.

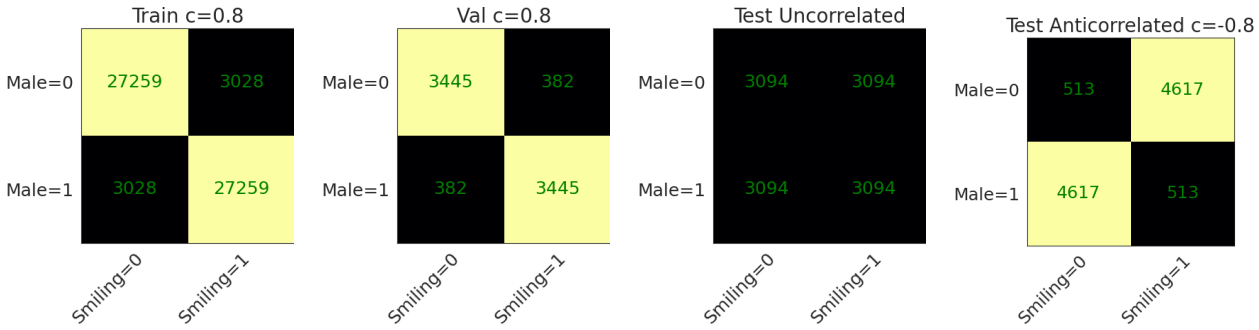
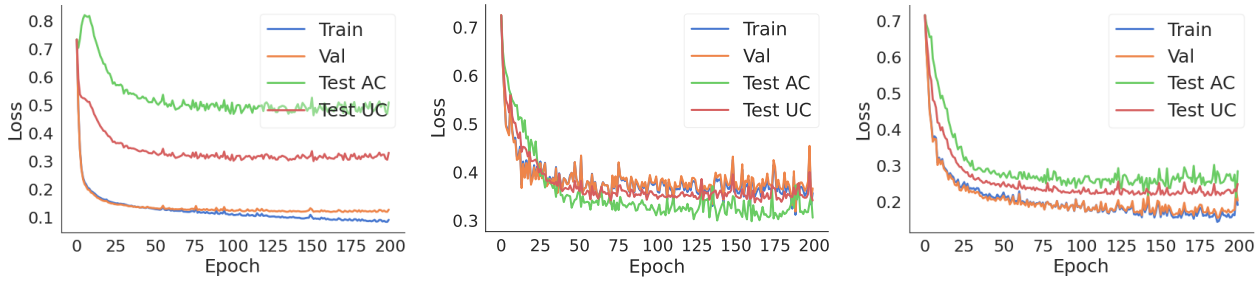
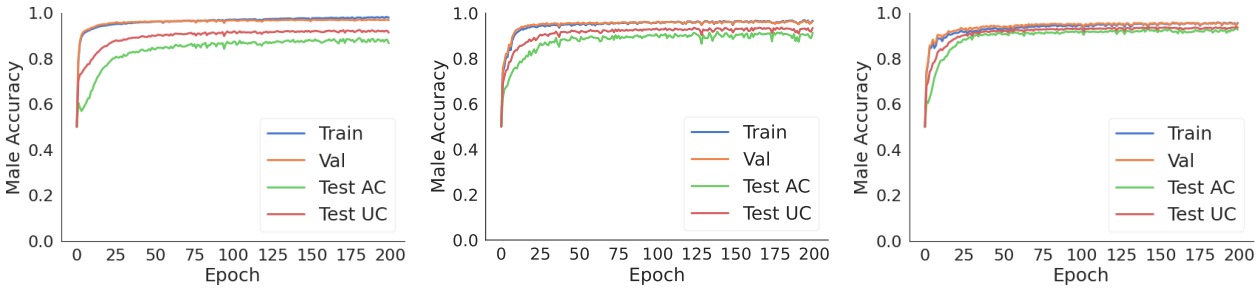


Figure 14: Numbers of examples in the subsampled CelebA datasets for the strongest correlation we consider, $c = 0.8$.



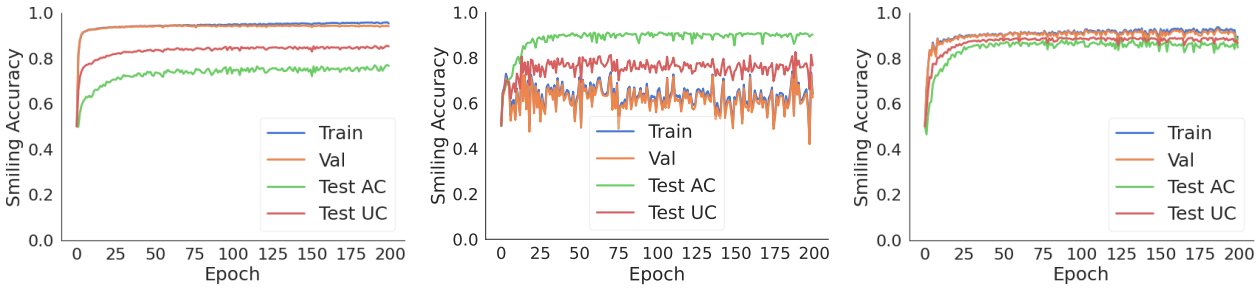
(a) Base (b) Base + MI (c) Base + CMI

Figure 15: Loss curves for each approach on the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.



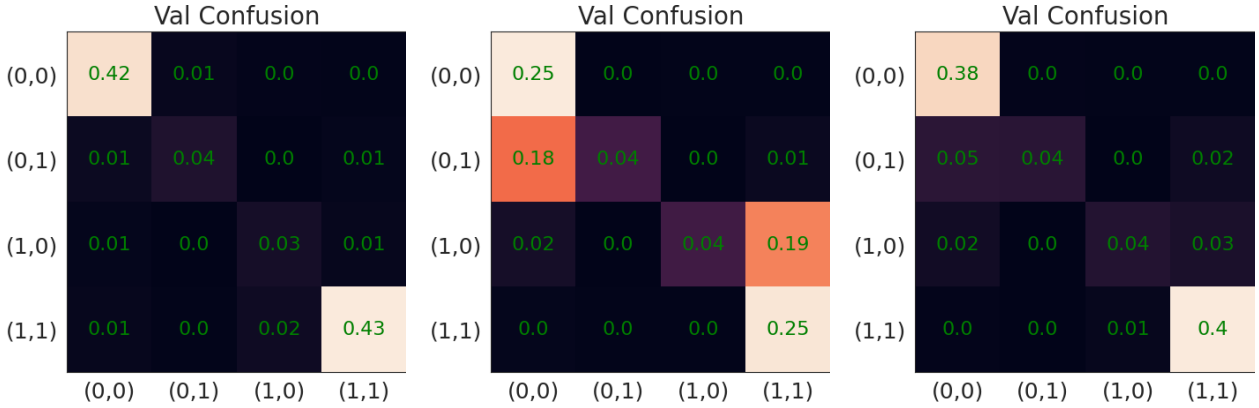
(a) Base (b) Base + MI (c) Base + CMI

Figure 16: Accuracies on the attribute `Male` for each approach on the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.



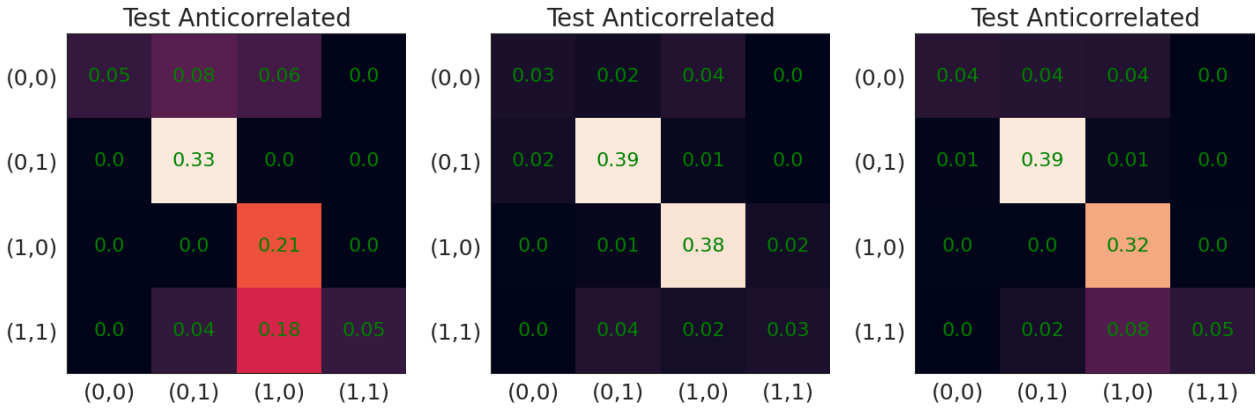
(a) Base (b) Base + MI (c) Baes + CMI

Figure 17: Accuracies on the attribute `Smiling` for each approach on the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.



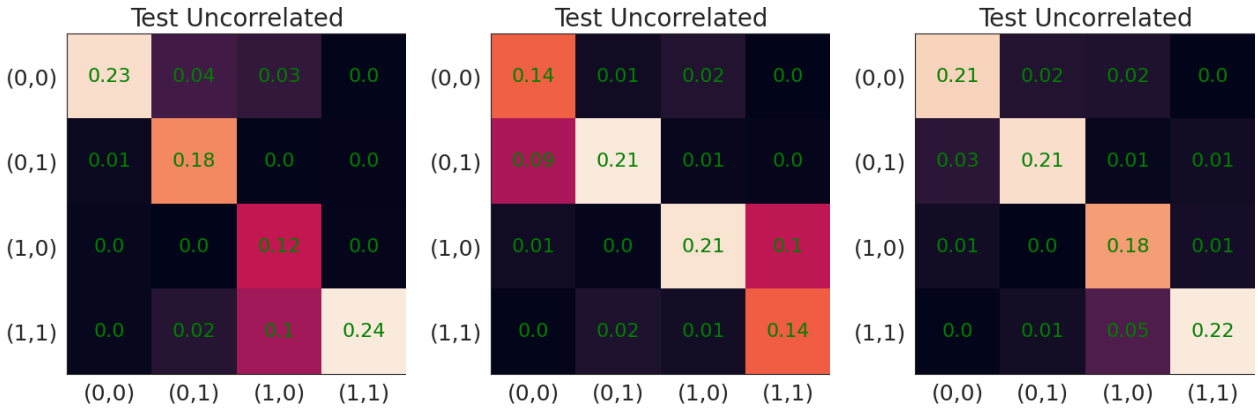
(a) Base (b) Base + MI (c) Base + CMI

Figure 18: Confusion matrices for each approach on the correlated validation set of the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.



(a) Base (b) Base + MI (c) Base + CMI

Figure 19: Confusion matrices for each approach on the anti-correlated test set of the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.



(a) Base (b) Base + MI (c) Base + CMI

Figure 20: Confusion matrices for each approach on the uncorrelated test set of the `Male-Smiling` CelebA task, under the strongest correlation we consider, $c = 0.8$.

	Common Combinations		**Rare Combinations**	
	Female + Non-Smiling	Male + Smiling	Female + Smiling	Male + Non-Smiling
Base	**4%**	**4%**	29%	51%
Base + MI	23%	28%	**12%**	31%
Base + CMI	10%	9%	20%	**29%**

Table 3: Percentage of incorrect predictions per subgroup for CelebA, evaluated on natural data (e.g., data with naturally-occurring correlations, that has not been subsampled to induce a specific correlation strength), using models trained on correlated data with $c = 0.8$.

	Common Combinations		**Rare Combinations**	
	Female + Non-Smiling	Male + Smiling	Female + Smiling	Male + Non-Smiling
Base	**4%**	**5%**	33%	49%
Base + MI	24%	28%	**11%**	26%
Base + CMI	9%	9%	19%	**25%**

Table 4: Percentage of incorrect predictions per subgroup for CelebA, evaluated on validation data ($c = 0.8$), using models trained on correlated data with $c = 0.8$.

B.4 DISENTANGLEMENT METRICS

We evaluated common disentanglement metrics (Locatello et al., 2019b) on uncorrelated test data using models trained on correlated data. We performed this analysis for two of our datasets and found in both cases that *Base+CMI* reached better scores compared to the other objectives for almost all metrics.

Toy Classification: Disentanglement results for the toy classification task with ten attributes are shown in Table 5. We obtained similar results for two and four attributes, which are not reported for brevity.

CelebA: Since the disentanglement metrics require that the factors of variation are each encoded in one-dimensional subspaces, we set latent dimension $D = 2$ for this experiment. In Table 6, we report the average and 68% confidence intervals for five models trained on data with correlation level 0.8.

Metric	**Base**	**Base+MI**	**Base+CMI**
IRS (Suter et al., 2019) ↑	0.377	0.573	**0.605**
SAP (Kumar et al., 2017) ↑	0.118	0.470	**0.477**
MIG (Chen et al., 2018) ↑	0.179	0.939	**0.975**
DCI Disentanglement (Eastwood & Williams, 2018) ↑	0.413	0.980	**0.998**
Beta-VAE (Higgins et al., 2017a) ↑	0.996	1	1
Factor-VAE (Kim & Mnih, 2018) ↑	1	1	1
Gaussian Total Correlation ↓	10.073	0.485	**0.025**
Gaussian Wasserstein Corr ↓	12.905	0.373	**0.027**
Gaussian Wasserstein Corr Norm ↓	0.866	0.037	**0.002**
Mutual Info Score ↓	0.975	0.197	**0.149**

Table 5: **Disentanglement metrics for toy classification with ten attributes.** Metrics are evaluated on the uncorrelated test set. Bold font indicates model with best disentanglement score.

Metric	**Base**	**Base+MI**	**Base+CMI**
IRS ↑	0.524 ± 0.043	$\mathbf{0.548 \pm 0.038}$	0.531 ± 0.041
SAP ↑	0.306 ± 0.003	0.296 ± 0.046	$\mathbf{0.389 \pm 0.005}$
MIG ↑	0.506 ± 0.01	0.455 ± 0.074	$\mathbf{0.674 \pm 0.007}$
DCI Disentanglement ↑	0.46 ± 0.009	0.596 ± 0.038	$\mathbf{0.807 \pm 0.023}$
Beta-VAE ↑	1.0 ± 0.0	1.0 ± 0.0	$\mathbf{1.0 \pm 0.0}$
Factor-VAE ↑	1.0 ± 0.0	0.999 ± 0.003	$\mathbf{1.0 \pm 0.0}$
Gaussian Total Correlation ↓	0.222 ± 0.012	0.056 ± 0.061	$\mathbf{0.011 \pm 0.003}$
Gaussian Wasserstein Corr ↓	0.351 ± 0.039	0.01 ± 0.009	$\mathbf{0.002 \pm 0.001}$
Gaussian Wasserstein Corr Norm ↓	0.098 ± 0.005	0.006 ± 0.004	$\mathbf{0.005 \pm 0.001}$
Mutual Info Score ↓	0.302 ± 0.022	0.111 ± 0.052	$\mathbf{0.042 \pm 0.006}$

Table 6: **Disentanglement metrics for CelebA.** Metrics are evaluated on the uncorrelated test set. Bold font indicates model with best disentanglement score.

B.5 WEAKLY SUPERVISED SETTING

For the fully supervised CelebA experiment, labels for both attributes were available for all 10260 images. For the weakly supervised setting, we reduced the number of labels to 5130 (50% of the labels of the fully supervised dataset), 2565 (25%), 1026 (10%), or 513 (5%) for each attribute. This implies that some images had both labels, some had only one label and some images had no labels (for example when using 50% of the labels the distinction is as follows: 25% of the images had both labels; 25% had only labels for attribute 1; 25% had only labels for attribute 2; and 25% had no labels). The three objectives can be applied to these weakly supervised settings. For *Base*, the cross-entropy loss for each attribute was computed only for the images that had labels for the corresponding attribute. For *Base+MI* no labels are required for the unconditional shuffling; thus this objective can be applied even for the images without labels. For *Base+CMI*, our method shuffles only images that have the same value for a given attribute. This also works if the labels of the other attribute are missing. We used the same training parameters as for the supervised experiment, except for increasing the number of training epochs (up to 1200 epochs) and adapting the minibatch size to the number of labels. In Figure 8 we report the average and 68% confidence intervals over three runs with different seeds.

C ALGORITHMS

In this section, we provide formal descriptions of the baseline approaches we use. Algorithm 2 describes the classification-only baseline, that trains separate linear classifiers to predict attributes $\mathbf{s}_k$ from the corresponding latent subspaces $\mathbf{z}_k$. Algorithm 3 and Algorithm 4 describe the unconditional disentanglement baseline, that adversarially minimizes the discrepancy between samples from the joint distribution $p(\mathbf{z}_1, \ldots, \mathbf{z}_k)$ and the product of marginals $p(\mathbf{z}_1) \cdots p(\mathbf{z}_k)$. Algorithm 5 describes the discriminator training loop for the CMI minimization approach from Section 4.

Algorithm 2 Supervised Learning on Subspaces

Input: $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}$, initial parameters for K linear classifiers $R_1, \ldots, R_K$
Input: $\boldsymbol{\theta}$, initial parameters for the encoder f
Input: α, β learning rates for training the encoder and linear classifiers
while true **do**
  $(\mathbf{x}, \{\mathbf{s}_k\}_{k=1}^K) \sim \mathcal{D}_{\text{Train}}$ ▷ Sample a minibatch of data with attribute labels
  $\mathbf{z} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{x})$ ▷ Forward pass through the encoder
  $\{\mathbf{z}_k\}_{k=1}^K \leftarrow \text{SplitSubspaces}(\mathbf{z}, k)$ ▷ Partition the latent space into k subspaces
  $L \leftarrow \sum_{k=1}^K L_{\text{cls}}(R_k(\mathbf{z}_k; \boldsymbol{\phi}_k), \mathbf{s}_k)$ ▷ Cross-entropy for each attribute
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} L$ ▷ Update encoder parameters
  $\boldsymbol{\phi}_k \leftarrow \boldsymbol{\phi}_k - \beta \nabla_{\boldsymbol{\phi}_k} L \quad , \quad \forall k \in \{1, \ldots, K\}$ ▷ Update classifier parameters
end while

Algorithm 3 Learning Unconditionally Disentangled Subspaces — Training the Encoder

Input: $\{\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_K\}$, initial parameters for K linear classifiers $R_1, \ldots, R_K$
Input: $\boldsymbol{\theta}$, initial parameters for the encoder f
Input: α, β learning rates for training the encoder and linear classifiers
while true **do**
  $(\mathbf{x}, \{\mathbf{s}_k\}_{k=1}^K) \sim \mathcal{D}_{\text{Train}}$ ▷ Sample a minibatch of data with attribute labels
  $\mathbf{z} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{x})$ ▷ Forward pass through the encoder
  $\{\mathbf{z}_k\}_{k=1}^K \leftarrow \text{SplitSubspaces}(\mathbf{z}, k)$ ▷ Partition the latent space into k subspaces
  $L \leftarrow \sum_{k=1}^K L_{\text{cls}}(R_k(\mathbf{z}_k; \boldsymbol{\phi}_k), \mathbf{s}_k)$ ▷ Cross-entropy for each attribute
  $\mathbf{z}' \sim p(\mathbf{z}_1)p(\mathbf{z}_2) \cdots p(\mathbf{z}_k)$ ▷ Samples w/ batchwise-shuffled subspaces
  $L \leftarrow L + \log\left(1 - D_{\boldsymbol{\omega}}(\mathbf{z}')\right) + \log\left(D_{\boldsymbol{\omega}}(\mathbf{z})\right)$ ▷ Add adversarial loss
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla_{\boldsymbol{\theta}} L$ ▷ Update encoder parameters
  $\boldsymbol{\phi}_k \leftarrow \boldsymbol{\phi}_k - \beta \nabla_{\boldsymbol{\phi}_k} L \quad , \quad \forall k \in \{1, \ldots, K\}$ ▷ Update classifier parameters
end while

Algorithm 4 Learning Unconditionally Disentangled Subspaces — Training the Discriminator

Input: $\boldsymbol{\omega}$, initial parameters for the discriminator D
Input: γ, learning rate for training the discriminator
while true **do**
  $(\mathbf{x}, \{\mathbf{s}_k\}_{k=1}^K) \sim \mathcal{D}_{\text{Train}}$ ▷ Sample a minibatch of data with attribute labels
  $\mathbf{z} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{x})$ ▷ Forward pass through the encoder
  $\{\mathbf{z}_k\}_{k=1}^K \leftarrow \text{SplitSubspaces}(\mathbf{z}, k)$ ▷ Partition the latent space into k subspaces
  $\mathbf{z}' \sim p(\mathbf{z}_1)p(\mathbf{z}_2) \cdots p(\mathbf{z}_k)$ ▷ Samples w/ batchwise-shuffled subspaces
  $L \leftarrow L + \log\left(D_{\boldsymbol{\omega}}(\mathbf{z}')\right) + \log\left(1 - D_{\boldsymbol{\omega}}(\mathbf{z})\right)$ ▷ Add adversarial loss
  $\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} - \gamma \nabla_{\boldsymbol{\omega}} L$ ▷ Update discriminator parameters
end while

Algorithm 5 Learning Conditionally Disentangled Subspaces Adversarially – Training the Discriminator

Input: $\boldsymbol{\omega}$, initial parameters for the discriminator D
Input: γ, learning rate for training the discriminator
while true **do**
  $(\mathbf{x}, \{\mathbf{s}_k\}_{k=1}^K) \sim \mathcal{D}_{\text{Train}}$ ▷ Sample a minibatch of data with attribute labels
  $\mathbf{z} \leftarrow f_{\boldsymbol{\theta}}(\mathbf{x})$ ▷ Forward pass through the encoder
  $\{\mathbf{z}_k\}_{k=1}^K \leftarrow \text{SplitSubspaces}(\mathbf{z}, k)$ ▷ Partition the latent space into K subspaces
  $L \leftarrow 0$ ▷ L will accumulate the losses over all subspaces
  for $k \in \{1, \ldots, K\}$ **do**
    $\mathbf{z}' \sim p(\mathbf{z}_1, \ldots \mathbf{z}_K \mid \mathbf{s}_k)$ ▷ Samples from the joint distribution
    $\mathbf{z}'' \sim p(\mathbf{z}_k \mid \mathbf{s}_k)p(\mathbf{z}_{-k} \mid \mathbf{s}_k)$ ▷ Samples w/ batchwise-shuffled subspaces
    $L \leftarrow L + \log\left(D_{\boldsymbol{\omega}}(\mathbf{z}'')\right) + \log\left(1 - D_{\boldsymbol{\omega}}(\mathbf{z}')\right)$ ▷ Add adversarial loss
  end for
  $\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} - \gamma \nabla_{\boldsymbol{\omega}} L$ ▷ Update discriminator parameters
end while

D PROOF OF PROPOSITION 3.1

Proposition 3.1 *If $I(s_1; s_2) > 0$, then enforcing $I(z_1; z_2) = 0$ means that $I(z_k; s_k) < H(s_k)$ for at least one k.*

Proof. Assume that $I(s_1; s_2) > 0$ and at the same time $I(z_k; s_k) = H(s_k)$ (i.e., we are proving by contradiction). Since $I(z_1; s_1) = H(s_1)$, we have $H(s_1 \mid z_1) = 0$ and with $H(s_1 \mid z_1) = H(s_1 \mid z_1, s_2) + I(s_1; s_2 \mid z_1)$ (both non-negative), it follows that $H(s_1 \mid z_1, s_2) = I(s_1; s_2 \mid z_1) = 0$. Since for the interaction information, by definition $I(s_1; s_2; z_1) = I(s_1; s_2) - I(s_1; s_2 \mid z_1)$, and $I(s_1; s_2 \mid z_1) = 0$, we have $I(s_1; s_2; z_1) = I(s_1; s_2) > 0$. Since we also assume $H(s_2 \mid z_2) = 0$, we also have $I(s_1; s_2; z_2) = I(s_1; s_2) > 0$.

We can use this to compute the fourth order interaction information $I(s_1; s_2; z_1; z_2)$. By definition, we have $I(s_1; s_2; z_1; z_2) = I(s_1; s_2; z_1) - I(s_1; s_2; z_1 \mid z_2)$. We just showed that $I(s_1; s_2; z_1) = I(s_1; s_2)$, and therefore we have $I(s_1; s_2; z_1 \mid z_2) = I(s_1; s_2 \mid z_2)$. Together it follows that:

$$
\begin{aligned}
I(s_1; s_2; z_1; z_2) &= I(s_1; s_2; z_1) - I(s_1; s_2; z_1 \mid z_2) && (7)\\
&= I(s_1; s_2) - I(s_1; s_2 \mid z_2) && (8)\\
&= I(s_1; s_2; z_2) && (9)\\
&= I(s_1; s_2) > 0 && (10)
\end{aligned}
$$

On the other hand, we know that $0 = H(s_1 \mid z_1) = H(s_1 \mid z_1; z_2) + I(s_1, z_2 \mid z_1)$ and therefore $I(s_1, z_2 \mid z_1) = 0$. Therefore, the interaction information $I(s_1; z_2; z_1) = I(s_1; z_2) - I(s_1; z_2 \mid z_1) = I(s_1; z_2) \geq 0$. At the same time, we assumed that $I(z_1; z_2) = 0$ and hence $I(z_1; z_2; s_1) + I(z_1; z_2 \mid s_1) = 0$, which shows that $I(z_1; z_2; s_1) \leq 0$. Together, we see that $I(z_1; z_2; s_1) = I(s_1; z_2) = 0$.

Now we can decompose $I(s_1; s_2; z_1; z_2)$ in a different way: $I(s_1; s_2; z_1; z_2) = I(s_1; z_1; z_2) - I(s_1; z_1; z_2 \mid s_2)$. We know that $I(s_1; z_1; z_2) = I(s_1; z_2)$ and therefore $I(s_1; z_1; z_2 \mid s_2) = I(s_1; z_2 \mid s_2) > 0$ and that $I(s_1; z_1; z_2) = 0$. Therefore, it follows that:

$$
\begin{aligned}
I(s_1; s_2; z_1; z_2) &= I(s_1; z_1; z_2) - I(s_1; z_1; z_2 \mid s_2) && (11)\\
&= 0 - I(s_1; z_2 \mid s_2) && (12)\\
&\leq 0 && (13)
\end{aligned}
$$

which is a contradiction with $I(s_1; s_2; z_1; z_2) = I(s_1; s_2) > 0$. Therefore, if $I(s_1; s_2) > 0$ and $I(z_1; z_2) = 0$, it must hold that $I(z_k; s_k) < H(s_k)$ for at least one k, which we wanted to show. □