

Contrastive Variational Autoencoder Enhances Salient Features

Abubakar Abid¹ James Zou²

Abstract

Variational autoencoders are powerful algorithms for identifying dominant latent structure in a single dataset. In many applications, however, we are interested in modeling latent structure and variation that are enriched in a target dataset compared to some background—e.g. enriched in patients compared to the general population. Contrastive learning is a principled framework to capture such enriched variation between the target and background, but state-of-the-art contrastive methods are limited to linear models. In this paper, we introduce the contrastive variational autoencoder (cVAE), which combines the benefits of contrastive learning with the power of deep generative models. The cVAE is designed to identify and enhance salient latent features. The cVAE is trained on two related but unpaired datasets, one of which has minimal contribution from the salient latent features. The cVAE explicitly models latent features that are shared between the datasets, as well as those that are enriched in one dataset relative to the other, which allows the algorithm to isolate and enhance the salient latent features. The algorithm is straightforward to implement, has a similar run-time to the standard VAE, and is robust to noise and dataset purity. We conduct experiments across diverse types of data, including gene expression and facial images, showing that the cVAE effectively uncovers latent structure that is salient in a particular analysis.

1. Introduction

In standard unsupervised learning, e.g. principal component analysis (PCA), the goal is to identify subspaces or latent variables that capture the dominant variation in one dataset of interest. In many applications, however, we have multiple datasets and we are most interested in identifying patterns or

¹Department of Electrical Engineering, Stanford University, California, United States ²Department of Biomedical Data Science, Stanford University, California, United States. Correspondence to: James Zou <jamesz@stanford.edu>.

variations that are *enriched* in a target dataset compared to some background data. For example, in biomedical analysis, we are often interested in patterns that are specific to data from the treatment group that are not shared with the control group. Popular unsupervised learning methods work on each dataset separately and do not try to capture variation enriched in a particular dataset when contrasted against some background.

Contrastive learning algorithms provide a way to identify such patterns. For example, contrastive principal components analysis (cPCA) is a recently-proposed method that seeks to identify *salient* principal components in a target dataset by identifying those linear combinations of features that are enriched in that dataset relative to a background dataset, rather than those that simply have the most variance (Abid et al., 2018). The cPCA algorithm has been shown to discover latent factors in data that are missed by standard unsupervised learning methods including PCA, and has been extended to time-series decomposition (Dirie et al., 2018). However, cPCA and its extensions have so far been limited to identifying linear patterns and subspaces within a dataset.

In many complex datasets, such as those consisting of natural images or data collected from genomic assays, the key relationships between latent features and observed samples are highly nonlinear. Here, we develop the contrastive variational autoencoder (cVAE) to identify nonlinear latent variables that are enriched in one dataset relative to another.

Motivating example: diverse dermatology images. Consider a dataset of images of skin lesions that are collected from a diverse population in order to study disease-related variations in skin. Such a dataset will include images from individuals of different ethnicities and sexes. Training a standard variational autoencoder (VAE) and examining the dominant latent factors will likely identify features of the skin related to the demographic diversity of the subjects (e.g. skin color, and other features related to age and gender). More subtle features related to disease type and progression, which are of interest, may not appear as prominent latent factors, or may be entangled with the dominant latent factors. If we also have skin images from different, healthy participants, which contain image variations related to the demographics of the subjects but not relevant to diseases, we may be able to isolate these the salient disease latent factors by contrasting the disease and healthy image datasets.

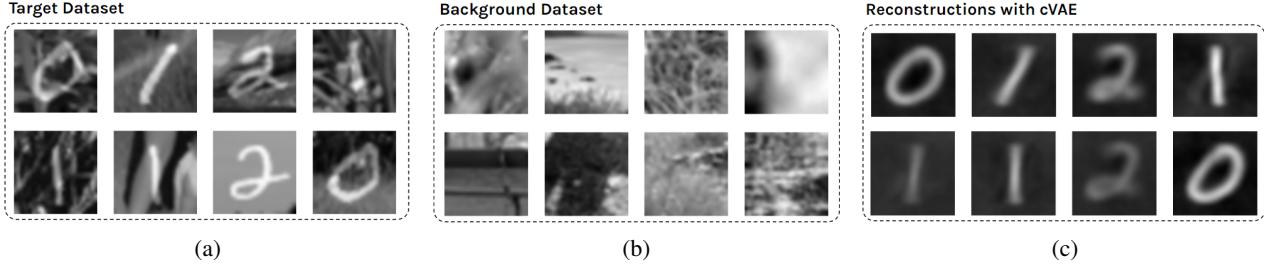


Figure 1. Contrastive variational autoencoders applied to the Grassy-MNIST dataset. (a) We construct a semi-synthetic dataset of hand-written digits from MNIST dataset superimposed on images of grass from the ImageNet dataset (b) We assemble a separate background dataset of grass images, also from the ImageNet dataset, but not ones used in the creation of the target dataset. (c) We train a cVAE on the target and background datasets. The cVAE identifies that the salient latent variables are those related to the digits. We can remove the irrelevant latent variables to generate clean hand-written digits, despite the network never being trained on such images. See Appendix A for reconstructions with standard VAEs

The cVAE algorithm that we develop in this paper offers a principled approach to do so. Note that the goal here is not to classify disease from healthy; such discriminatory learning approach often miss the complex variation within the disease group that we want to model (Abid et al., 2018).

As a generative model, the cVAE can further be used to generate new samples with varying levels of the salient and irrelevant latent variables. For example, in Fig. 1, we show how a cVAE trained on images of hand-written digits on complex grassy backgrounds and as well as a background dataset of only grass images can be used to generate brand new images of clean hand-written digits, despite the cVAE never observing such images during training. We test cVAEs on several kinds of datasets, and find that, with a suitable background dataset, they consistently outperform standard VAEs in identifying and isolating latent features that are salient to the analyst. We have made the code for our experiments are publicly available¹.

1.1. Related Work

Variational autoencoders (VAEs) are a powerful deep generative framework to capture latent structure in complex data (Kingma & Welling, 2013; Doersch, 2016). Much recent work has focused on learning disentangled representations with VAEs (Kim & Mnih, 2018; Chen et al., 2018), in which each latent feature learns one semantically meaningful factor of variation, while being invariant to other factors. Other work has focused on learning relationships *between* latent features, such as (Sønderby et al., 2016), which proposed learning latent features along with a dependency structure between them encoded as hierarchical network.

In distinguishing between salient and irrelevant latent fea-

tures, we adapt the principles of contrastive analysis that have been proposed in the context of unsupervised learning (Zou et al., 2013; Ge & Zou, 2016). These principles are utilized by the contrastive principal components analysis (cPCA) algorithm Abid et al. (2018) to distinguish between salient and irrelevant principal components by introducing a background dataset without the salient variance. Dirie et al. (2018) extended this idea to time series analysis by developing contrastive singular spectrum analysis (cSSA) to decompose a time series into salient subsignals that are absent from a secondary time series. Extending the ideas of contrastive analysis to VAEs was independently explored in concurrent work by (Severson et al., 2018), who has a more restricted formulation than the method proposed in this paper.

Our work also bears some resemblance to recent work in domain adaptation, where the goal is to learn domain-specific representation separately from factors that are shared across domains (Bousmalis et al., 2016; Gonzalez-Garcia et al., 2018). However, these methods are not designed to extract salient latent features in one dataset in an unsupervised manner; the method proposed by (Bousmalis et al., 2016) is designed with intention of feeding shared features into a single classifier, while Gonzalez-Garcia et al. (2018) primarily design their method, based on a pair of generative adversarial networks (GANs), as a means for sampling and image translation. Because the goals of these methods is different than ours, the architecture and loss functions are also considerably more complex, with a large number of components that are not needed for the cVAE algorithm.

2. Problem Setting

We begin by describing the generative latent variable model that motivates the development of the cVAE algorithm. Consider the samples $\{\mathbf{x}_i\}_{i=1}^n$ in our target dataset to be n i.i.d.

¹https://github.com/abidlabs/contrastive_vae

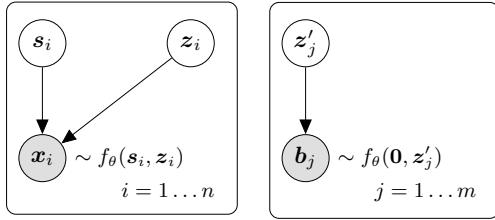


Figure 2. Generative model for target and background. The generative model for each sample of the target data \mathbf{x}_i and background data \mathbf{b}_j , includes two sets of latent variables, $\mathbf{s}, \mathbf{z} \sim \mathcal{N}(0, I)$. The former consists of those latent variables that are salient to the analyst, and the latter consists of those variables which are irrelevant. Here, $\mathbf{x}_i \sim f_\theta(\mathbf{s}_i, \mathbf{z}_i)$. For the background dataset, the presence of the salient latent variables is minimal or absent, so we have $\mathbf{b}_j \sim f_\theta(\mathbf{0}, \mathbf{z}'_j)$. Observed quantities are shaded.

samples that are generated by a random process that involves two sets of latent variables $\mathbf{z}_i, \mathbf{s}_i$, which are drawn from a known prior distribution. The observed samples are drawn from a conditional distribution that is an arbitrary function f_θ of the latent variables. This function is parametrized by unknown parameters θ :

$$\mathbf{x}_i \sim f_\theta(\mathbf{x}|\mathbf{z}_i, \mathbf{s}_i) \quad (1)$$

We use \mathbf{s} to refer to the latent variables that are *salient* in a particular analysis. In the example presented in the introduction, these may be the latent variables related to subtypes of skin cancer, or the progression of the disease. The second set of variables \mathbf{z} are the latent variables *irrelevant* for the analysis, such as features related to demographic variations among the individuals.

Trained on such a dataset, a standard VAE will not necessarily discover the salient latent features. In fact, if the irrelevant latent features generate most of the variation in the data, the VAE will identify those irrelevant features as they allow the VAE to minimize the reconstruction loss more effectively. Alternatively, the VAE may learn latent representations in which the irrelevant variables are entangled with the relevant latent variables.

How can we identify and isolate the salient latent variables? Following the contrastive analysis framework, we introduce a secondary background dataset $\{\mathbf{b}_j\}_{j=1}^m$ in which the salient latent features are absent:

$$\mathbf{b}_j \sim f_\theta(\mathbf{b}|\mathbf{0}, \mathbf{z}'_j). \quad (2)$$

Let us assume that the unobserved random variables $\mathbf{s}_i, \mathbf{z}_i, \mathbf{z}_j$ are independently drawn from the isotropic Gaussian prior, and here, we use different indices i and j to indicate the the target and background datasets are *unpaired*,

though the samples in each dataset come from the same generative process $f_\theta(\cdot)$. The generative model is shown as a probabilistic graphical model in Fig. 2.

Background datasets naturally exist in many settings. For example, we have observed that in datasets collected from diseased patients, where the salient latent factors are those related to disease type and progression, a control group of healthy subjects can provide a suitable background. More generally, when population-level variations are to be excluded, one can often use samples drawn from a similar population as the target dataset, but without the variation of interest. Conversely, a homogeneous sample can be used as the background for a sample drawn from a mixed population to eliminate the intra-group variation and measurement noise, while enhancing the inter-group variation. In other settings, such as the Grassy-MNIST dataset introduced in Fig. 1, a set of signal-free data can serve as a good background to remove structured noise from samples that include both signal and noise.

The learning problem that we are faced with then becomes designing an algorithm for learning the parameters θ of the transformation $f_\theta(\cdot)$ from unpaired target and background datasets. We are also interested in approximate posterior inference of the latent features. In particular, we seek to distinguish the salient latent features \mathbf{s} and the irrelevant latent features \mathbf{z} .

3. Method

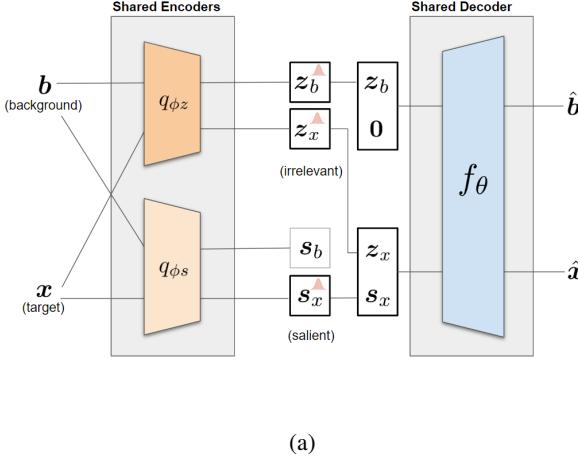
To approach this problem, we introduce two probabilistic encoders, denoted by $q_{\phi_s}(\mathbf{s}|\mathbf{x})$ and $q_{\phi_z}(\mathbf{z}|\mathbf{x})$, which are parametrized approximations to the intractable posteriors over the latent variables \mathbf{s} and \mathbf{z} respectively. We can derive a variational bound on the likelihood of individual data points, in terms of the unknown parameters, similar to that of the standard variational autoencoder (Kingma & Welling, 2013). For the target dataset, we have the following likelihood lower bound (see Appendix B for derivation):

$$\mathcal{L}_x(\mathbf{x}_i) \geq \mathbb{E}_{q_{\phi_s}(\mathbf{s})q_{\phi_z}(\mathbf{z})} [f_\theta(\mathbf{x}_i|\mathbf{s}, \mathbf{z})] - \text{KL}(q_{\phi_s}(\mathbf{s}|\mathbf{x}_i)\|p(\mathbf{s})) - \text{KL}(q_{\phi_z}(\mathbf{z}|\mathbf{x}_i)\|p(\mathbf{z})) \quad (3)$$

For the likelihood of the background dataset, we have:

$$\mathcal{L}_b(\mathbf{b}_i) \geq \mathbb{E}_{q_{\phi_z}(\mathbf{z})} [f_\theta(\mathbf{x}_i|\mathbf{0}, \mathbf{z})] - \text{KL}(q_{\phi_z}(\mathbf{z}|\mathbf{b}_i)\|p(\mathbf{z})) \quad (4)$$

Now, we train neural networks to maximize the sum of objective functions (3) and (4) by training two encoder networks q_{ϕ_s} and q_{ϕ_z} , to infer \mathbf{s} and \mathbf{z} respectively from the observed features, and a single decoder network $f_\theta(\cdot)$ which takes as input a concatenated version of \mathbf{x} and \mathbf{z} as input and reconstructs the original samples. Such an architecture



Algorithm 1. Training a Contrastive Variational Autoencoder

Input: target dataset $\{\mathbf{x}_i\}_{i=1}^n$, background dataset $\{\mathbf{b}_j\}_{j=1}^m$, encoder network q_ϕ , decoder network f_θ , discriminator D_ψ , # of steps S , batch size B , learning rate λ

Create two instances of the encoder network: q_{ϕ_x} and q_{ϕ_z}

Initialize the parameters of the encoders, decoder, and discriminator.

for step $\in \{1 \dots S\}$ **do**

- Randomly select batches of target & background data $\{\mathbf{x}^{(i)}\}, \{\mathbf{b}^{(i)}\}$ of size B
- Sample $\mathbf{s}_x^{(i)} \sim q_{\phi_x}(\mathbf{x}^{(i)})$, $\mathbf{z}_x^{(i)} \sim q_{\phi_z}(\mathbf{x}^{(i)})$ for $i = 1 \dots B$
- Sample $\mathbf{s}_b^{(i)} \sim q_{\phi_x}(\mathbf{b}^{(i)})$, $\mathbf{z}_b^{(i)} \sim q_{\phi_z}(\mathbf{b}^{(i)})$ for $i = 1 \dots B$
- Compute $\mathcal{L}_x(\mathbf{x}^{(i)})$ and $\mathcal{L}_b(\mathbf{b}^{(i)})$ as defined in (3) and (4) respectively
- Let $\mathbf{v}^{(i)}$ be a single vector that is the concatenation of $\mathbf{s}_x^{(i)}$ and $\mathbf{z}_x^{(i)}$
- Approximate TC⁽ⁱ⁾ as $\log(D_\psi(\mathbf{v}^{(i)})/1 - D_\psi(\mathbf{v}^{(i)}))$.
- Define the objective $G = \sum_i^B \mathcal{L}_x(\mathbf{x}^{(i)}) + \mathcal{L}_b(\mathbf{b}^{(i)}) + TC^{(i)}$
- Update $\theta \leftarrow \theta + \lambda \nabla_\theta G$, $\phi_s \leftarrow \phi_s + \lambda \nabla_{\phi_s} G$, $\phi_z \leftarrow \phi_z + \lambda \nabla_{\phi_z} G$
- Let $\tilde{\mathbf{v}}^{(i)}$ be $\mathbf{v}^{(i)}$ with the $\mathbf{s}_x^{(i)}$ and $\mathbf{z}_x^{(i)}$ columns randomly swapped w.p. 1/2
- Define the discriminator objective $G' = \sum_i^B \log D_\psi(\mathbf{v}^{(i)})(1 - D_\psi(\tilde{\mathbf{v}}^{(i)}))$
- Update $\psi \leftarrow \psi + \lambda \nabla_\psi G'$

end for

Return: trained encoders q_{ϕ_x}, q_{ϕ_z} and decoder f_θ

(b)

Figure 3. Architecture and algorithm for the cVAE. (a) Here, we illustrate the architecture of the contrastive VAE. The background target datasets are fed into two sets of shared encoders q_{ϕ_s} and q_{ϕ_z} , which output the salient and irrelevant latent variables respectively. The latent variables for the target dataset are concatenated and fed into a decoder network, while only the (zero-padded) irrelevant variable is fed into the decoder network. The decoder outputs reconstructions of the target and background data. (b) The pseudocode for the cVAE is provided here, using batch gradient descent for simplicity.

is shown in Fig. 3(a). This allows to use stochastic gradient descent on mini-batches of the target and background dataset to learn the parameters of the three neural networks.

Empirically, we observe that we can further improve results by encouraging independence between salient and irrelevant latent features. We add a *total correlation* (TC) term to the objective function for the target dataset,

$$TC = -KL(\bar{q} \| q_{\phi_s}(s|\mathbf{x}_i) \cdot q_{\phi_z}(z|\mathbf{x}_i)), \quad (5)$$

where $\bar{q} \stackrel{\text{def}}{=} q_{\phi_s, \phi_z}(s, z|\mathbf{x}_i)$ is the joint conditional probability of s, z . If s and z are independent, then (5) is zero; otherwise, it is negative. However, calculating these terms over mini-batches of samples is difficult to do directly. To approximate this term, we use the density-ratio trick, analogous to the method described in Kim & Mnih (2018). In essence, we use a separate discriminator to distinguish samples $[s, z]$ drawn from \bar{q} and from $q_{\phi_s}(s|\mathbf{x}_i) \cdot q_{\phi_z}(z|\mathbf{x}_i)$. The classification probabilities of this discriminator can be used to estimate (5). This discriminator is trained simultaneously with the encoder and decoder neural networks described earlier. More details are provided in Appendix C.

Once the parameters ϕ_s , ϕ_z , and θ are learned, the cVAE can be used to solve three related problems that make it particularly useful for settings in which there are both salient and irrelevant latent features:

1. **Infer only the salient features** by feeding a sample \mathbf{x}_i into q_{ϕ_s} and sampling from the resulting distribution.

2. **Generate samples using only the salient features** by sweeping values of s_i , concatenating s_i with $\mathbf{0}$, and feeding the result into the decoder f_θ .
3. **Denoise samples (i.e. remove irrelevant structure)** by feeding a sample \mathbf{x}_i into q_{ϕ_s} and sampling s_i from the resulting distribution, then concatenating s_i with $\mathbf{0}$, and feeding the result through the decoder f_θ .

For the latter two tasks, notice that we have set $z_i = \mathbf{0}$ to simulate the effect of removing the irrelevant latent features. For linear encoders and decoders (i.e. PCA), this would remove any influence of the irrelevant latent variables. However, for neural networks, this is not generally the case, since a neural network may transform an input of $\mathbf{0}$ to a non-zero output. To ensure that this does not happen, in experiments where we denoise samples, we constrain all of the neural networks q_{ϕ_s} , q_{ϕ_z} , and f_θ to have zero biases. Coupled with ReLU activations, this ensures that an input of $\mathbf{0}$ to any of these networks does not result in a non-zero output. Pseudocode to train cVAEs is shown in Fig. 3(b).

4. Evaluation

We evaluate cVAEs on several target and background datasets, both quantitatively and qualitatively. On the quantitative side, we choose datasets in which there exist ground-truth classes, but which are not detected easily using only the dominant latent features. We train both standard VAEs and cVAEs and evaluate whether the samples, when pro-

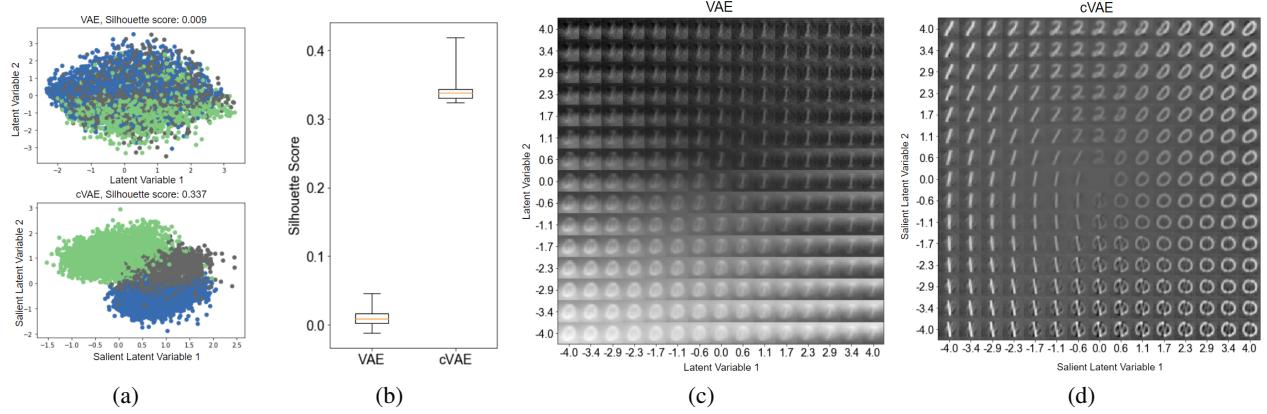


Figure 4. Evaluation of the VAE and cVAE on the Grassy-MNIST dataset. Here, we show the results of applying both the standard VAE and cVAE to the Grassy-MNIST dataset, introduced in Fig. 1. (a) The top panel shows that when the samples are embedded into the 2-dimensional latent space of the VAE, they do not cluster by digit (here, images with the digit 0 are in green, with the digit 1 are in blue, and the digit 2 in gray). The bottom panel shows that when the samples are embedded into the 2-dimensional *s*-latent space of the VAE, they do cluster by digit. (b) These results are also consistent, as shown in this boxplot which shows the resulting silhouette scores across 10 trials. (c) Here, we use the trained VAE to generate new samples by sweeping values in the 2-dimensional latent space. The generated samples include both digit features and background features. (d) Here, we use the trained cVAE to generate new samples by sweeping values in the 2-dimensional latent space resulting in clean hand-written digits. In Appendix D, we show that even a standard VAE with a larger latent space does not learn digit-related factors in this setting.

jected into the learned latent space, produce distinct clusters in agreement with classes that present in the target dataset. We use this method to evaluate because it represents the unsupervised learning regime, in which an analyst without class labels may visualize or embed points in a lower dimension using VAEs to reveal salient substructure.

We use the *silhouette score* (SS) to quantify the degree to which the clustering agrees with ground-truth class labels (Rousseeuw, 1987). This metric has been used in prior literature to assess the quality of latent VAE embeddings (Hu & Greene, 2018; Clachar, 2016). The SS is defined as follows: for a single data point i , let $SS(i) \stackrel{\text{def}}{=} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$, where $a(i)$ is the average distance from i to the other points that share its label, while $b(i)$ is the average distance from i to the points in the next nearest cluster. The SS for a dataset is the mean $SS(i)$ for all of the points i in the dataset, and varies between -1 and 1 , with a higher positive value indicating that clusters agree more strongly with class labels.

On a qualitative level, we have shown in Fig. 1 that the cVAE is able to denoise images, removing complex grassy backgrounds present in a target dataset of hand-written digits. We further show that the cVAE can be used to generate samples in which only the salient features vary, making the cVAE useful as a way to precisely adjust salient features independently of irrelevant ones when generating samples.

4.1. Target and Background Datasets

Grassy-MNIST. We constructed this semi-synthetic target dataset by superimposing hand-written digits from the popular MNIST dataset (LeCun et al., 2010) with images labeled as grass in the ImageNet (Deng et al., 2009) dataset. Each image of grass was cropped, resized to be 28 by 28 pixels, converted to grayscale, and scaled to have double the amplitude of the MNIST digits before being superimposed on the digit. Only the digits 0, 1, and 2 were used in the target dataset. The background dataset consisted of other images of grass not used in the construction of the target dataset.

RNA-Seq with Batch Effects. We combined two single-cell RNA-Seq datasets released by Zheng et al. (2017) measuring the expression levels of genes from blood cells collected from a leukemia patient before and after transplant. The samples were labeled by which step (before/after the procedure) the cells were collected. These cells contained significant batch effects, so as a background, we selected a dataset of healthy blood cells released by the same authors and processed on the same equipment.

CelebA. We used the popular celebrity faces dataset (Liu et al., 2015) after cropping each image and scaling it to be 64 by 64 pixels. We used subsets of the CelebA dataset with certain attributes as the target. Namely, we formed one target dataset of celebrities wearing glasses, and we formed a second target dataset of celebrities wearing caps. In both cases, the remaining celebrities constituted the background.

4.2. Implementation Details

The VAE and cVAE models were implemented in Tensorflow², using the Keras API. The target and background datasets were fed simultaneously into the model. The loss function consisted of the sum of the negative of the terms in (3) and (4), along with the TC loss, and the loss of a discriminator used to estimate the TC (see Appendix C for more details). We used logistic regression as the discriminator.

We used the Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$. We used a batch size of 128 for our experiments. For the Grassy-MNIST and RNA Seq datasets, the VAE architecture consisted of fully-connected layers in the encoder and decoder. For the CelebA dataset, a convolutional VAE architecture was used. The exact architecture of our models can be found in Appendix E.

4.3. Quantitative Results

We first trained a standard VAE on the Grassy-MNIST target dataset with a latent dimensionality of 2. We embedded each sample in the latent space, producing the plot shown in Fig. 4(a, top), in which each sample is colored by the digit that appears in the image. In this latent space, the samples do not cluster by the digit (SS=0.009); this is because the dominant features are those that relate to the grassy background, rather than features of the hand-drawn digit.

We then trained a cVAE on the Grassy-MNIST target and background datasets with a s -latent dimension of 2 and an z -latent dimension of 2. We embedded each point in the dataset onto the 2D s -latent space, producing the plot shown in Fig. 4(a, bottom). We find that in this latent space, samples clustered strongly by digits (SS=0.337), indicating that the cVAE has identified digit-related features as the salient latent variables. This pattern is consistently reproducible; the results of 10 trials are shown in Fig. 4(b).

Next, we trained a standard VAE on the RNA-Seq dataset with a latent dimensionality of 2. We embedded each gene expression vector in the latent space, producing the plot shown in Fig. 5(a, top), in which each sample is colored to indicate whether it was pre/post-transplant. In this latent space, the samples do not cluster by the digit (SS=0.012); this is likely because the dominant features are those that relate to the batch effects or variations within cell types.

We then trained a cVAE on the RNA-Seq target and background datasets with a s -latent dimension of 2 and an z -latent dimension of 4. We embedded each point in the dataset onto the 2D s -latent space, producing the plot shown in Fig. 5(a, bottom). We find that in this latent space, samples clustered more strongly by pre/post-transplant status

(SS=0.227), indicating that the cVAE was able to better identify latent features corresponding to transplant status. Aggregate results over 10 trials are shown in Fig. 5(b).

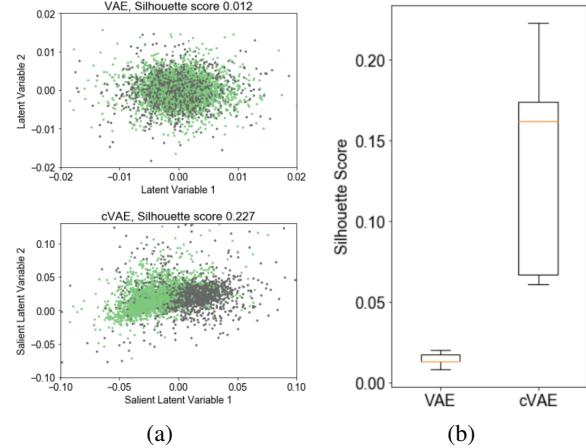


Figure 5. Comparing VAE and cVAE on the RNA-Seq dataset. Here, we show the results of applying both the standard VAE and cVAE to the RNA-Seq dataset. (a) The top panel shows that when the samples are embedded into the 2-dimensional latent space of the VAE, they do not cluster by transplant (here, pre-transplant samples are in green and post-transplant samples in gray). The bottom panel shows that when the samples are embedded into the 2-dimensional s -latent space of the cVAE, they do cluster by transplant status. A few points are omitted for ease of visualization; see complete plots and further analysis in Appendix F. (b) Aggregated silhouette scores over 10 trials are shown here.

4.4. Qualitative Results

Using the trained models on the Grassy-MNIST dataset, we generated new digits using the method described in Section 3. We show the results using the standard VAE and cVAE in Fig. 4(c) and Fig. 4(d) respectively. We find that the standard VAE generates new images with some degree of background structure. This background structure is entangled with digit features, with the digit ‘0’ containing more background than the digit ‘1’. Furthermore, the digit ‘2’ is not generated at all, as the features needed to reconstruct the digit are not learned by the standard VAE. On the other hand, we notice by using a contrastive VAE, we are able to generate the clean digits ‘0’, ‘1’, and ‘2’, despite the network never being trained on such images.

We then determine whether we are able to observe similar results on a more complex dataset, the CelebA dataset. In this case, we use two attributes, the presence of eyeglasses and the presence cap, to define our two target datasets. We use the remaining images as the background datasets. The purpose is to extract latent features related to the target attribute. For example, if we are interested in extracting features related to eyeglasses shape or color or size, we

²<https://www.tensorflow.org/>

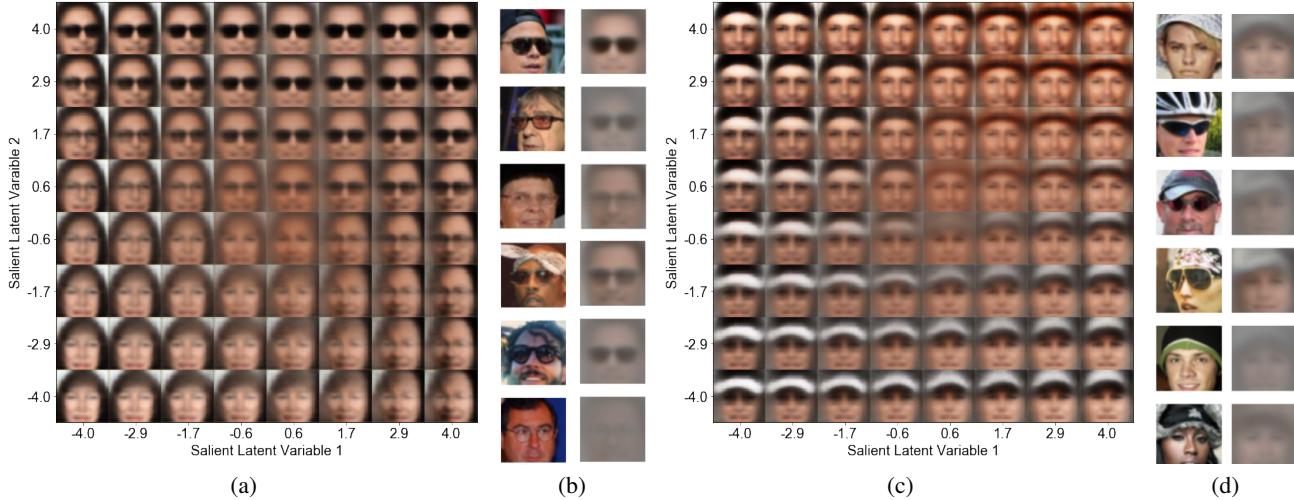


Figure 6. Results of applying a cVAE to the CelebA Dataset. Here, we apply cVAEs to a target dataset consisting of celebrity images either wearing glasses (for panels a-b) or caps (for panels c-d). In both cases, the remaining celebrities constituted the background dataset. (a) We use the trained cVAE to generate new samples of celebrity images by varying the two learned latent factors. We see that the two latent factors concern attributes of the glasses. In particular, latent variable 1 seems to deal with glasses direction, while latent variable 2 is related to the transparency/presence of the glasses. (b) If we denoise the images by only selecting the salient features learned by the cVAE, we see that primarily the pixels that remain in the reconstructed image relate to the glasses. (c) In this example, the cVAE has learned that the salient latent features that relate to attributes of the cap, namely the color of the cap and its position. (d) If select only the salient features learned by the cVAE, we see that the pixels that relate to the cap in the original image are primarily present in the reconstruction, although some facial features are present as well. (*cf.* figures in Appendix A, which show the results using standard VAEs.)

would define a target dataset of images of celebrities with glasses, and a background dataset in which none of the celebrities wear glasses.

We carry out such an analysis in Fig. 6. In Fig. 6(a), we find that the cVAE extracts salient features related to the eyeglasses, and in Fig. 6(c), the cVAE extracts features related to the caps worn by the celebrities. Compared to results using standard VAEs (see Appendix A), we find that the cVAE more clearly extracts features that are related to the target attribute, and these are less entangled with dominant latent variables in the image, such as skin color and background color. This is further supported by experiments shown in Fig. 6(b) and Fig. 6(d), in which we use the trained cVAEs to “denoise” the images, removing the effects of the irrelevant latent variables. We find, particularly in Fig. 6(b), that this erases most of the pixels in the image, leaving only the features needed to reconstruct to the target attribute e.g. the glasses or the cap.

4.5. Sensitivity Analyses

In order to understand the reliability of cVAEs in diverse settings, we measured the robustness of cVAEs to various factors, such as the construction of the target/background datasets as well as the dimensionality of the latent space. We carried out experiments using the semi-synthetic Grassy-

MNIST dataset. First, we varied the scale of the background images of grass used to construct the target images. We started with a scale of 0, representing no grassy images in the target data (i.e. the target was the original MNIST dataset) and, in increments of 0.25, increased the scale to 2 (i.e. the amplitude of the grassy background was twice that of the digits). The background dataset was unchanged.

At each increment, we compared the efficacy of a standard VAE to that of a cVAE in identifying salient latent features. The results are shown in Fig. 7(a). At low levels of background, both the standard VAE and the cVAE were effective at clustering MNIST digits. However, as the grassy background increased in amplitude, the efficacy of the standard VAE dropped significantly, while the cVAE continued to be able to discern relevant salient features.

We then considered how well-tailored a background dataset needed to be to serve as an effective contrast for a target dataset. We left the target dataset unchanged, but added varying levels of random isotropic noise to the background images of grass. We started with a noise level of 0, representing no noise in the background data (i.e. the background images were the original images of grass from the ImageNet dataset), and in increments of 0.25, increased the scale to 2, at which point the images of grass were completely obscured by noise. Nevertheless, the performance of the cVAE did not seem to suffer significantly, as shown in Fig. 7(b).

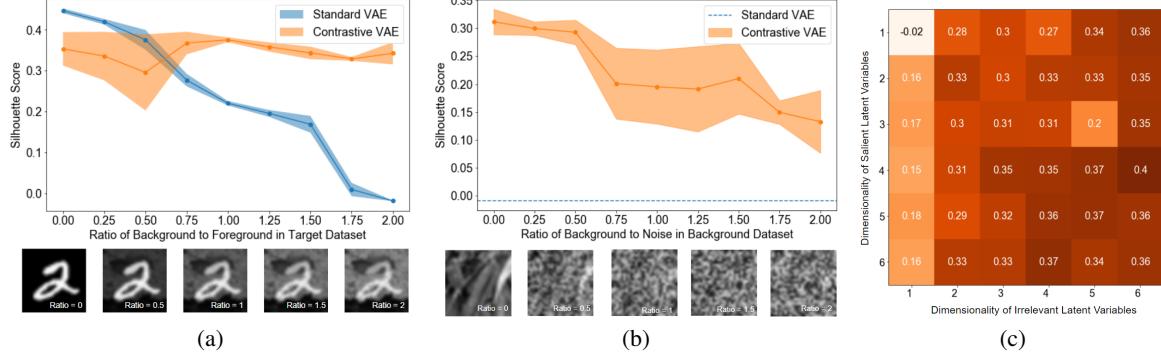


Figure 7. Sensitivity of the cVAE to Dataset Composition and Hyperparameters. Here, we probe the sensitivity of the cVAE algorithm to various factors. (a) We vary the construction of the Grassy-MNIST dataset by changing the relative amounts of ‘background’ (images of grass), compared to the ‘foreground’ (images of the digits) in each target image. We find that when there is little-to-no background, both kinds of VAEs perform well, with the standard VAE performing slightly better. With the introduction of the background, the performance of the standard VAE degrades, but not that of the contrastive VAE. Sample target images are shown below the plot. (b) Next, we add varying levels of isotropic noise to the background dataset, to see whether an imperfect background dataset affects performance. We do not observe significant reduction in performance even with the ratio of the amplitude of the noise-to-background increases. (c) Here, we explore the effect of changing the dimensionality of the s and z latent spaces. The Silhouette score for each combination of hyperparameters is listed in each spot, and the color scale reflects the amplitude of these scores. Further sensitivity analyses are included in Appendix G.

We also considered the importance of a hyperparameter in the cVAE algorithm: the dimensionality of the latent space used for s and z . We varied each of these from 1-6 and trained a cVAE with those dimensions on the Grassy-MNIST dataset. We recorded the resulting silhouette scores in Fig. 7(c). We found that increasing the size of the z -latent space consistently tended to improve results. Increasing the size of the s -latent space did not seem to improve or impair results. Additionally, we performed experiments to characterize the sensitivity of the cVAE to dataset composition; these experiments are described in Appendix G.

5. Discussion

In this paper, we have introduced the cVAE, an extension of the standard VAE algorithm that allows users to identify latent factors that are salient to their own analyses through the use of a background dataset. We discuss the methodology behind the cVAE in Section 3, where we derive it using a variational approximation to the maximum likelihood problem. In Section 4, we empirically compare the cVAE architecture to that of the standard VAE architecture on biological and image datasets, finding that the cVAE algorithm allows the user to: (1) identify salient latent factors, (2) generate new samples with only salient features, and (3) denoise samples by removing the irrelevant latent variables.

We have shown, for example, that the cVAE can be used to discover latent factors that allow digits to be classified even in the presence of complex noise (see Fig. 4) or cells to be categorized even with strong batch effects (Fig. 5). In a dataset of celebrity images, cVAEs were not completely

able to remove the effect of irrelevant latent variables, perhaps because they may be correlated with the salient latent variables in complex ways. However, they offered a way to enhance and more finely control the latent variables of interest (Fig. 6).

It is worth emphasizing that the cVAE architecture is not designed for *discriminative* analyses: the identified latent factors are not intended to discriminate between points in the target and background dataset. Rather, the cVAE architecture helps the user identify and isolate salient latent factors that may be otherwise not present amongst or entangled with the dominant latent factors in a target dataset.

The choice of the background dataset has a large influence on the result of cPCA. In general, the background data should have the structure or latent features that one would like to remove from the target data. Nevertheless, through systematic experiments, we find that the cVAE algorithm is fairly robust to noise in the background datasets, suggesting that even when the background dataset contains random latent factors that are not present in the target dataset, performance is not significantly adversely affected.

The primary strengths of the contrastive VAE is its generality and ease of use. Training a cVAE essentially requires the same runtime as a standard VAE and can be done in most situations where VAEs are commonly used today. Doing so allows users to shift the focus from finding those latent features that are dominant in data to those that are salient in the context of their analyses.

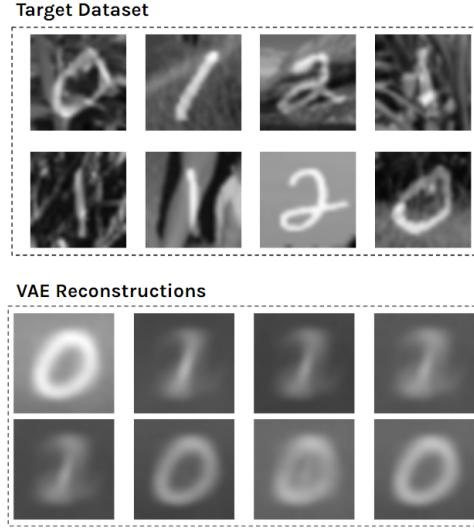
References

- Abid, A., Zhang, M. J., Bagaria, V. K., and Zou, J. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications*, 9(1):2134, 2018.
- Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. Domain separation networks. In *Advances in Neural Information Processing Systems*, pp. 343–351, 2016.
- Chen, T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.
- Clachar, S. *Novelty detection and cluster analysis in time series data using variational autoencoder feature maps*. The University of North Dakota, 2016.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 248–255. Ieee, 2009.
- Dirie, A.-H., Abid, A., and Zou, J. Contrastive multivariate singular spectrum analysis. *arXiv preprint arXiv:1810.13317*, 2018.
- Doersch, C. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.
- Ge, R. and Zou, J. Rich component analysis. In *ICML*, pp. 1502–1510, 2016.
- Gonzalez-Garcia, A., van de Weijer, J., and Bengio, Y. Image-to-image translation for cross-domain disentanglement. *arXiv preprint arXiv:1805.09730*, 2018.
- Hu, Q. and Greene, C. S. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell rna transcriptomics. *bioRxiv*, pp. 385534, 2018.
- Kim, H. and Mnih, A. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.
- Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.
- Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.
- Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.
- Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.
- Severson, K., Ghosh, S., and Ng, K. Unsupervised learning with contrastive latent variable models. *arXiv preprint arXiv:1811.06094*, 2018.
- Sønderby, C. K., Raiko, T., Maaløe, L., Sønderby, S. K., and Winther, O. Ladder variational autoencoders. In *Advances in neural information processing systems*, pp. 3738–3746, 2016.
- Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8:14049, 2017.
- Zou, J. Y., Hsu, D. J., Parkes, D. C., and Adams, R. P. Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems*, pp. 2238–2246, 2013.

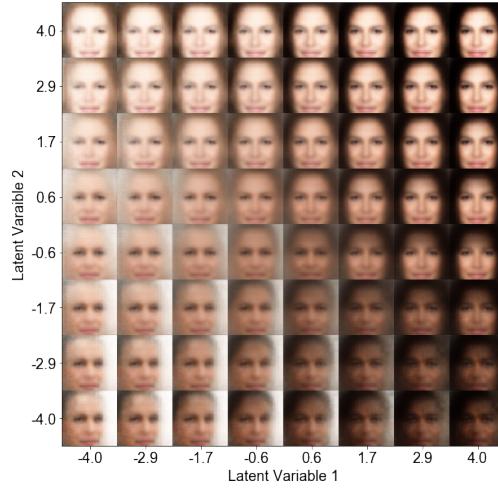
Appendices

A. Qualitative Results with Standard VAEs

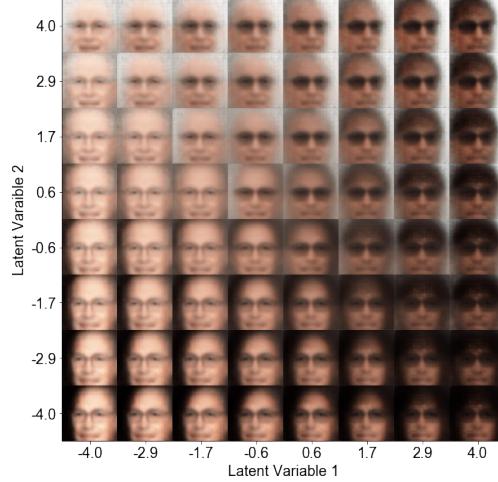
In Fig. 1, we showed how cVAE reconstructions from the Grassy-MNIST dataset could be used to generate new clean images of hand-written digits never seen by the generative model. For comparison, here we show the reconstructions of a standard VAE:



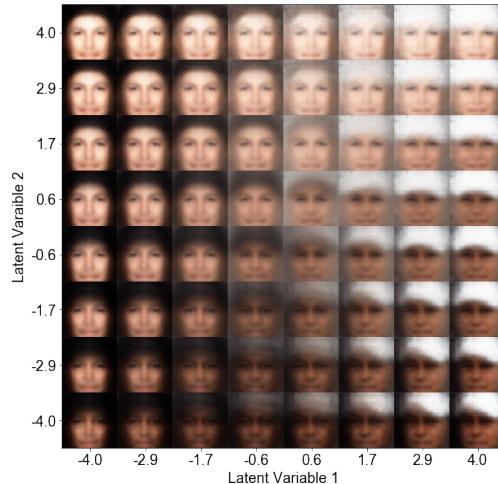
Next, we show the result of training a standard convolutional variational autoencoder with latent dimensionality of 2 on the CelebA dataset. Sweeping the latent variables produces the following images, which suggest that the dominant latent structure in the Celeb A dataset are related to skin color and background behind, or orientation of, the face.



In Fig. 6(a), we showed that we could isolate latent variables related to eyeglasses by using a contrastive VAE. For comparison, here we show the reconstructions of a standard VAE trained *only* on the target dataset of celebrities with glasses. Because we are restricting ourselves to celebrities with glasses, the presence/absence of glasses does appear to be a latent variable. However, this suffers in comparison to the cVAE in two ways: (1) the presence/absence of glasses factor is entangled with the dominant variation in the datasets (related to skin color and background behind the face), making it difficult to adjust in isolation. (2) the latent feature does not relate to variation *within* the population of glasses-wearing celebrities, but rather serves to differentiate between those who wear dark glasses and those who wear light/no glasses.



Similarly, in Fig. 6(c), we showed that we could isolate latent variables related to eyeglasses by using a contrastive VAE. For comparison, here we show the reconstructions of a standard VAE trained *only* on the target dataset of celebrities with hats. Because we are restricting ourselves to celebrities with hats, the presence/absence of hats does appear to be a latent variable. However, this suffers in comparison to the cVAE in two ways: (1) the presence/absence of hats factor is entangled with the dominant variation in the datasets (related to skin color and background behind the face), making it difficult to adjust in isolation. (2) the latent feature does not relate to variation *within* the population of hat-wearing celebrities, but rather serves to differentiate between those who wear white hats and those who wear dark/no hats.



B. Derivation of Variational Lowerbound

In this appendix, we derive the variational lower bounds presented in (3) and (4) in the main text. We start by writing the log-likelihood for each sample in the target as:

$$\log p_\theta(\mathbf{x}^{(i)}) = \log \sum_{\mathbf{s}, \mathbf{z}} p_\theta(\mathbf{x}^{(i)}, \mathbf{s}, \mathbf{z}) \quad (6)$$

$$= \log \sum_{\mathbf{s}, \mathbf{z}} \frac{p_\theta(\mathbf{x}^{(i)}, \mathbf{s}, \mathbf{z}) \cdot q_\phi(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z})}{q_\phi(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z})} \quad (7)$$

$$\geq \sum_{\mathbf{s}, \mathbf{z}} q_\phi(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z}) \log \frac{p_\theta(\mathbf{x}^{(i)}, \mathbf{s}, \mathbf{z})}{q_\phi(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z})} \quad (8)$$

$$= \sum_{\mathbf{s}, \mathbf{z}} q_\phi(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z}) \log \frac{p_\theta(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z}) \cdot p(\mathbf{s}, \mathbf{z})}{q_\phi(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z})} \quad (9)$$

$$= \sum_{\mathbf{s}, \mathbf{z}} q_\phi(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z}) \log p_\theta(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z}) + q_\phi(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z}) \log \frac{p(\mathbf{s}, \mathbf{z})}{q_\phi(\mathbf{x}^{(i)}, \mathbf{s}, \mathbf{z})} \quad (10)$$

$$= \mathbb{E}_{q_\phi(\mathbf{s}, \mathbf{z})} [\log p_\theta(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z})] - \text{KL}(q_\phi(\mathbf{s}, \mathbf{z} | \mathbf{x}^{(i)}) \| p(\mathbf{s}, \mathbf{z})) \quad (11)$$

$$= \mathbb{E}_{q_{\phi_s}(\mathbf{s}) q_{\phi_z}(\mathbf{z})} [\log p_\theta(\mathbf{x}^{(i)} | \mathbf{s}, \mathbf{z})] - \text{KL}(q_{\phi_s}(\mathbf{s} | \mathbf{x}^{(i)}) \| p(\mathbf{s})) - \text{KL}(q_{\phi_z}(\mathbf{z} | \mathbf{x}^{(i)}) \| p(\mathbf{z})) \quad (12)$$

where, in (8), we have used Jensen's inequality, and in (12), we have used the independence of \mathbf{z} and \mathbf{s} to decompose the joint KL divergence into the sum of two marginal KL divergence.

This gives us the variational bound in (3). The derivation for the log-likelihood of the background data, (4), is even simpler, as it consists of only one set of latent variables; it follows the same derivation as the standard variational autoencoder (Kingma & Welling, 2013).

C. Total Correlation Loss

In this appendix, we describe the total correlation loss in more detail, and how we compute it efficiently with contrastive VAEs. The total correlation loss is defined as:

$$\text{TC} = -\text{KL}(\bar{q} \| q_{\phi_s}(s|\mathbf{x}_i) \cdot q_{\phi_z}(z|\mathbf{x}_i)), \quad (13)$$

where $\bar{q} \stackrel{\text{def}}{=} q_{\phi_s, \phi_z}(s, z|\mathbf{x}_i)$. This quantity cannot be easily approximated during stochastic gradient descent, since its batch estimate suffers from computational issues (see [Kim & Mnih \(2018\)](#), Appendix D for more details). Instead, we use the density-ratio trick to estimate the total correlation.

The key idea behind the density-ratio trick ([Nguyen et al., 2010](#)) is to train a classifier that tries to predict whether a sample $[s, z]$ is drawn from \bar{q} or from $q_{\phi_s}(s|\mathbf{x}_i) \cdot q_{\phi_z}(z|\mathbf{x}_i)$. We simulate samples from the latter term by randomly shuffling the z and s dimensions.

The probabilities outputted by such a classifier can be used to estimate the desired KL term. Therefore, in our algorithm, we train a simple logistic regression classifier D_ψ along with the cVAE. This classifier is included in the pseudocode shown in Fig. 3(b). We compute the following quantity as the estimate of the total correlation:

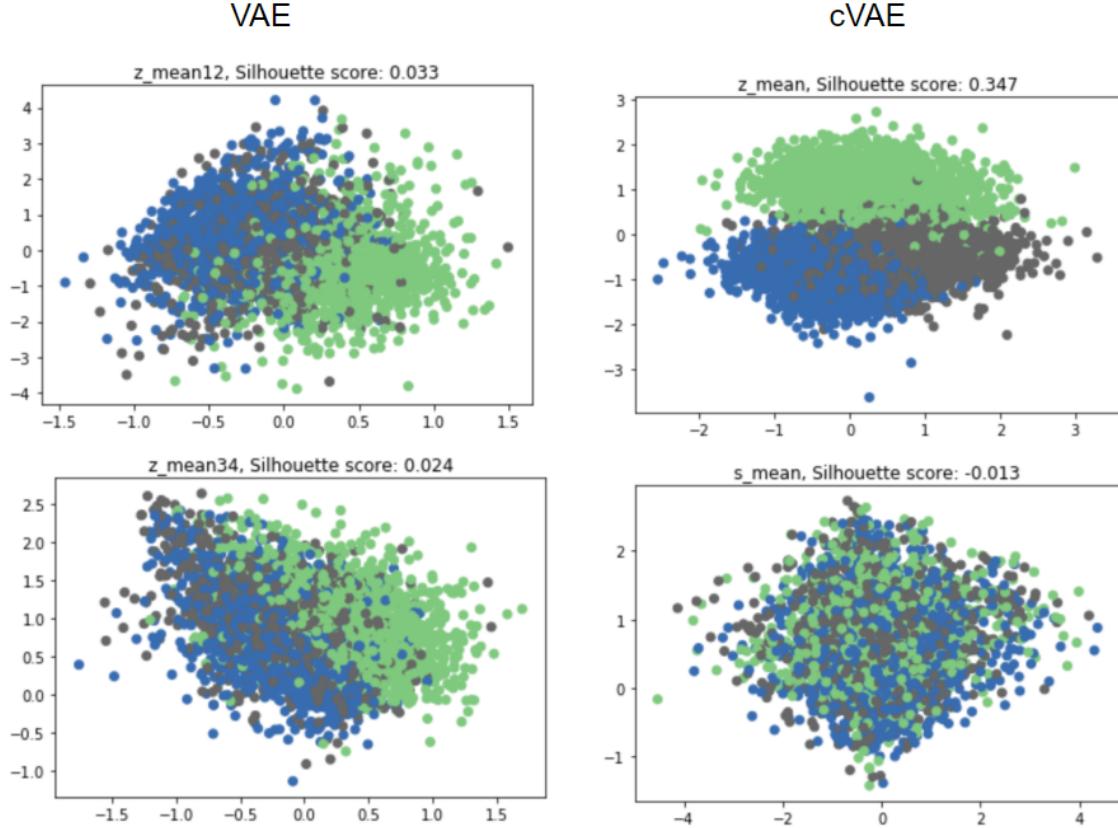
$$\text{TC} \approx \sum_i \log(D_\psi(\mathbf{v}^{(i)}) / 1 - D_\psi(\mathbf{v}^{(i)})), \quad (14)$$

where $\mathbf{v}^{(i)}$ is the concatenation of $[s^{(i)}, z^{(i)}]$, and the summation is carried over the batch of samples.

D. Additional Quantitative Results on the Grassy-MNIST Dataset

In this appendix, we include two sets of results. On the left, we train a standard VAE in four dimensions on the target dataset in Grassy-MNIST. We observe that even with four latent dimensions (the same total number of dimensions as the contrastive VAE), we still do not observe any latent dimensions which individually allow any kind of discrimination of the class digit. The top left plot shows the first and second latent dimensions, while the bottom left plot shows the third and fourth latent dimensions.

On the right, top, we show the same result as in Fig. 4(a, top), in which we show the salient latent variables learned by the cVAE can be used to discriminate between digits. Additionally, in the bottom right, we also show the irrelevant variables learned by the cVAE. As expected, these do not vary in significant way among the classes of digits.



E. Architecture Details

For the experiments in this paper, we use two kinds of architectures for the VAEs/cVAEs. For the experiments with the Grassy-MNIST dataset, as well as the RNA-Seq dataset, we use a fully-connected architecture. The second kind of architecture includes has convolutional layers in the encoder and transpose-convolutional layers in the decoder. This architecture is used for experiments with the the CelebA dataset. The layers in each architecture are written below.

Layers in the fully-connected VAE/cVAE:

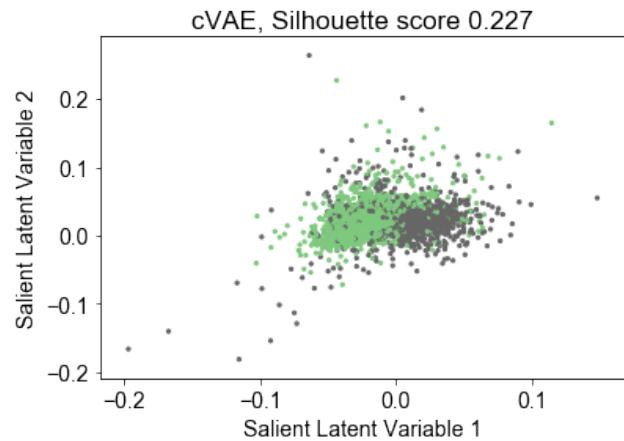
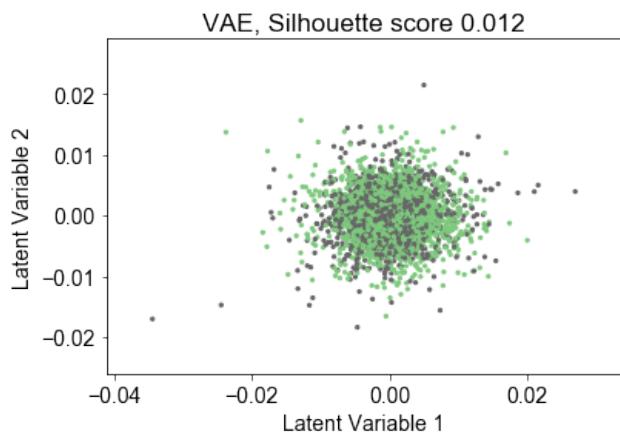
- input shape of 784 [for **Grassy-MNIST**] and 500 [for **RNA-Seq**]
- intermediate layer of dimensionality 128
- (salient) latent feature layer with dimensionality 2 (in the cVAE, there are irrelevant feature layers whose dimensionality varies from 1 to 6 depending on the experiment)
- intermediate layer of dimensionality 128
- output layer of the same dimensionality as the input layer

Layers in the convolutional VAE/cVAE:

- input shape of (64, 64, 3)
- 2D-convolutional layer with kernel size (5, 5) and a stride of (2, 2) with 32 channels
- 2D-convolutional layer with kernel size (5, 5) and a stride of (2, 2) with 64 channels
- intermediate fully-connected layer of dimensionality 128
- salient latent feature layer with dimensionality 2 (in the cVAE, there are irrelevant feature layers with dimensionality 4)
- intermediate fully-connected layer of dimensionality 128
- 2D-convolutional-transpose layer with kernel size (5, 5) and a stride of (2, 2) with 64 channels
- 2D-convolutional-transpose layer with kernel size (5, 5) and a stride of (2, 2) with 64 channels
- final output of (64, 64, 3)

F. Complete Results for Fig. 5

In Fig. 5(a), the plots we show omit a few points for ease of visualizations. Here, we show the complete plots.



G. Sensitivity Analyses with Respect to Dataset Composition

We conducted two additional experiments in order to characterize the sensitivity of the contrastive VAE to the *composition* of the target and background datasets. In the first experiment, we constructed a target dataset that included varying amounts of *background* images (of only grass). The background dataset was unchanged. We carried out the rest of the experiment as outlined in Section 4.5, in which we measured the resulting silhouette scores. As shown on the left, we found that the contrastive VAE performance degraded as more background images were introduced in the target dataset – which is not surprising, as the contrast between the target and background was weakened by the addition of background images to the target dataset.

In the second experiment, we did not change the target dataset, but instead changed the composition of the background dataset to include images from the target dataset. We measured the resulting silhouette scores. We found that the contrastive VAE performance surprisingly did not seem to degrade as more target images were introduced in the background dataset, as shown on the right.

