SepVAE: a contrastive VAE to separate pathological patterns from healthy ones

Robin Louiset [1 2] **Edouard Duchesnay** [1] **Antoine Grigis** [1] **Benoit Dufumier** [1 2] **Pietro Gori** [2]

Abstract

Contrastive Analysis VAE (CA-VAEs) is a family of Variational auto-encoders (VAEs) that aims at separating the common factors of variation between a *background* dataset (BG) (*i.e.,* healthy subjects) and a *target* dataset (TG) (*i.e.,* patients) from the ones that only exist in the target dataset. To do so, these methods separate the latent space into a set of **salient** features (*i.e.,* proper to the target dataset) and a set of **common** features (*i.e.,* exist in both datasets). Currently, all models fail to prevent the sharing of information between latent spaces effectively and to capture all salient factors of variation. To this end, we introduce two crucial regularization losses: a disentangling term between common and salient representations and a classification term between background and target samples in the salient space. We show a better performance than previous CA-VAEs methods on three medical applications and a natural images dataset (CelebA). Code and datasets are available on GitHub `https://github.com/neurospin-projects/2023_rlouiset_sepvae`.

1. Introduction

One of the goals of unsupervised learning is to learn a compact, latent representation of a dataset, capturing the underlying factors of variation. Furthermore, the estimated latent dimensions should describe distinct, noticeable, and semantically meaningful variations. One way to achieve that is to use a generative model, like Variational Auto-Encoders (VAEs) (Kingma & Welling, 2013), (Higgins et al., 2017) and disentangling methods (Higgins et al., 2017), (Burgess et al., 2018), (Shu et al., 2018),(Zheng & Sun, 2019), (Chen et al., 2019), (Ainsworth et al., 2018), (Li et al., 2018). Dif-

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA. 2023. Copyright 2023 by the author(s).

ferently from these methods, which use a *single* dataset, in Contrastive Analysis (CA), researchers attempt to distinguish the latent factors that generate a *target* (TG) and a *background* (BG) dataset. Usually, it is assumed that target samples comprise additional (or modified) patterns with respect to background data. The goal is thus to estimate the **common** generative factors and the ones that are **target-specific** (or **salient**). This means that background data are fully encoded by some generative factors that are also **common** with the target data. On the other hand, target samples are assumed to be partly generated from strictly proper factors of variability, which we entitle **target-specific** or **salient** factors of variability. This formulation is particularly useful in medical applications where clinicians are interested in separating common (i.e., healthy) patterns from the salient (i.e., pathological) ones in an *intepretable* way.

For instance, consider two sets of data: 1) healthy neuro-anatomical MRIs (BG=*background dataset*) and 2) Alzheimer-affected patients' MRIs (TG=*target dataset*). As in (Jack et al., 2018), (Antelmi et al., 2019), given these two datasets, neuroscientists would be interested in distinguishing common factors of variations (*e.g.:* effects of aging, education or gender) from Alzheimer's specific markers (*e.g.:* temporal lobe atrophy, an increase of beta-amyloid plaques). Until recently, separating the various latent mechanisms that drive neuro-anatomical variability in neuro-degenerative disorders was considered hardly feasible. This can be attributed to the intertwining between the variability due to natural aging and the variability due to neurodegenerative disease development. The combined effects of both processes make hardly interpretable the potential discovery of novel bio-markers.

The objective of developing such a Contrastive Analysis method would be to help separate these processes. And thus identifying correlations between neuro-biological markers and pathological symptoms. In the **common features** space, aging patterns should correlate with normal cognitive decline, while **salient features** (*i.e.:* Alzheimer-specific patterns) should correlate with pathological cognitive decline.

Besides medical imaging, Contrastive Analysis (CA) methods cover various kinds of applications, like in pharmacology (placebo versus medicated populations), biology (pre-intervention vs. post-intervention cohorts) (Zheng et al., 2017), and genetics (healthy vs. disease population (Jones
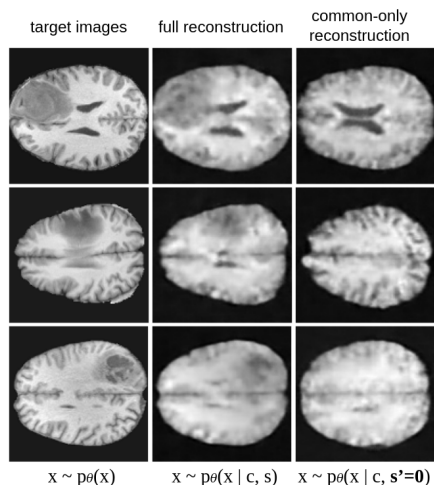
target images　　full reconstruction　　common-only reconstruction

$x \sim p_\theta(x)$　　$x \sim p_\theta(x \mid c, s)$　　$x \sim p_\theta(x \mid c, \mathbf{s'}=\mathbf{0})$

Figure 1. SepVAE reconstructions on Brats2021 dataset (Menze et al., 2015). (Middle) full reconstructions using the estimated common and salient latent vectors. (Right) common-only reconstructions using the estimated common latent vectors and fixing the salient factors to s'. The common latent variables encode the healthy factors of variability (*e.g.* : brain shape and aspect), while the salient factors encode the pathological patterns (*e.g.* : tumors), which are not visible in the right columns (common-only).

et al., 2021), (Haber et al., 2017)).

2. Related works

Variational Auto-Encoders (VAEs) (Kingma & Welling, 2013) have advanced the field of unsupervised learning by generating new samples and capturing the underlying structure of the data onto a lower-dimensional data manifold. Compared to linear methods (e.g., PCA, ICA), VAEs make use of deep non-linear encoders to capture non-linear relationships in the data, leading to better performance on a variety of tasks.

Disentangling methods (Higgins et al., 2017; Burgess et al., 2018; Shu et al., 2018) enable learning the underlying factors of variation in the data. While disentangling (Zheng & Sun, 2019; Chen et al., 2019) is a desirable property for improving the control of the image generation process and the interpretation of the latent space (Ainsworth et al., 2018; Li et al., 2018), these methods are usually based on a *single* dataset, and they do not explicitly use labels or multiple datasets to effectively estimate and separate the common and salient factors of variation.

Semi and weakly-supervised VAEs (Mathieu et al., 2019; Kingma et al.; Maaløe et al., 2016; Joy et al., 2021) have proposed to integrate class labels in their training. However, these methods solely allow conditional generalization and better semantic expressivity rather than addressing the separation of the factors of variation between distinct datasets.

Contrastive Analysis (CA) works are explicitly designed to identify patterns that are unique to a target dataset compared to a background dataset. First attempts (Zou et al., 2013; Abid et al., 2018; Ge & Zou, 2016) employed linear methods in order to identify a projection that captures the variance of the target dataset while minimizing the background information expressivity. However, due to their linearity, these methods had reduced learning expressivity and were also unable to produce satisfactory generation.

Contrastive VAE (Abid & Zou, 2019; Weinberger et al., 2022; Severson et al., 2019; Ruiz et al., 2019; Zou et al., 2022; Choudhuri et al., 2019) have employed deep encoders in order to capture higher-level semantics. They usually rely on a latent space split into two parts, a common and a salient, produced by two different encoders. First methods, such as (Severson et al., 2019), employed two decoders (common and salient) and directly sum the common and salient reconstructions in the input space. This seems to be a very strong assumption, probably wrong when working with high-dimensional and complex images. For this reason, subsequent works used a single decoder, which takes as input the concatenation of both latent spaces. Importantly, when seeking to reconstruct background inputs, the decoder is fed with the concatenation of the common part and an informationless reference vector **s'**. This is usually chosen to be a null vector in order to reconstruct a null (i.e., empty) image by setting the decoder's biases to 0. Furthermore, to fully enforce the constraints and assumptions of the underlying CA generative model, previous methods have proposed different regularizations. Here, we analyze the most important ones with their advantages and shortcomings:

Minimizing background's variance in the salient space Pioneer works (Severson et al., 2019; Abid & Zou, 2019) have shown inconsistency between the encoding and the decoding task. While background samples are reconstructed from **s'**, the salient encoder does not encourage the background salient latents to be equal to **s'**. To fix this inconsistency, posterior works (Weinberger et al., 2022; Zou et al., 2022; Choudhuri et al., 2019) have shown that explicitly nullifying the background variance in the salient space was beneficial. This regularization is necessary to avoid salient features explaining the background variability but not sufficient to prevent information leakage between common and salient spaces, as shown in (Weinberger et al., 2022).

Independence between common and salient spaces Only one work (Abid & Zou, 2019) proposed to prevent information leakage between the common and salient space by minimizing the total correlation (TC) between $q_{\phi_c, \phi_s}(c, s|x)$ and $q_{\phi_c}(c|x) \times q_{\phi_s}(s|x)$, in the same fashion as in Factor-VAE (Kim & Mnih, 2019). This requires to independently train a discriminator $D_\lambda(.)$ that aims at approximating the ratio between the joint distribution $q(x) = q_{\phi_c, \phi_s}(c, s|x)$ and
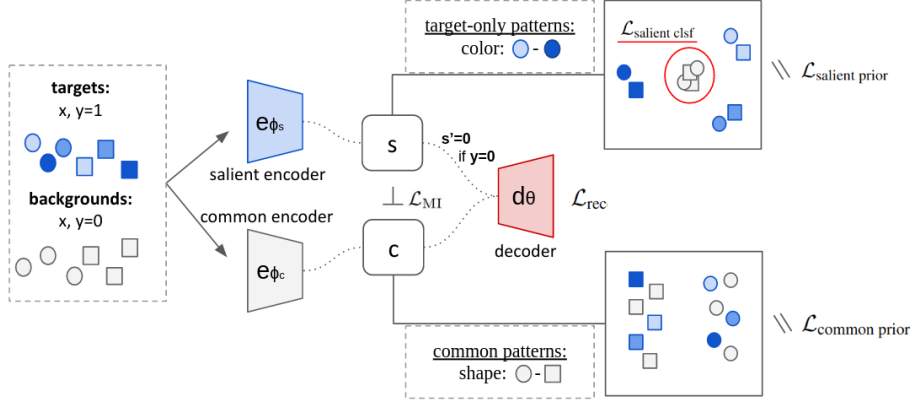
Figure 2. Illustration of SepVAE training. Target and background images are encoded with the same encoders e_{ϕ_s} and e_{ϕ_c}. The first encoder e_{ϕ_s} estimates the salient factors of variation s of the target samples ($y = 1$). Background samples ($y = 0$) salient space is set to an informationless value $s' = 0$. The second encoder e_{ϕ_c} estimates the common factors c. Images are reconstructed using a single decoder d_θ fed with the concatenation of c and s.

the marginal of the posteriors $\bar{q}(x) = q_{\phi_c}(c|x) \times q_{\phi_s}(s|x)$ via the density-ratio trick (Nguyen et al., 2010; Sugiyama et al., 2012). In practice, (Abid & Zou, 2019)'s code does not use an independent optimizer for λ, which undermines the original contribution. Moreover, when incorrectly estimated, the TC can become negative, and its minimization can be harmful to the model's training.

Matching background and target common patterns Another work (Weinberger et al., 2022), has proposed to encourage the distribution in the common space to be the same across target samples and background samples. Mathematically, it is equivalent to minimizing the KL between $q_{\phi_c}(c|y = 0)$ and $q_{\phi_c}(c|y = 1)$ (or between $q_{\phi_c}(c)$ and $q_{\phi_c}(c|y)$). In practice, we argue that it may encourage undesirable *biases* to be captured by salient factors rather than common factors. For example, let's suppose that we have healthy subjects (*background* dataset) and patients (*target* dataset) and that patients are composed of both young and old individuals, whereas healthy subjects are only old. We would expect the CA method to capture the normal aging patterns (*i.e.:* the bias) in the common space. However, forcing both $q_{\phi_c}(c|x, y = 0)$ and $q_{\phi_c}(c|x, y = 1)$ to follow the same distribution in the common space would probably bring to a biased distribution and thus to leakage of information between salient and common factors (i.e., aging could be considered as a salient factor of the patient dataset).This behavior is not desirable, and we believe that the statistical independence between common and salient space is a more robust property.

Contributions Our contributions are three-fold:
• We develop a new Contrastive Analysis method: SepVAE, which is supported by a sound and versatile Evidence Lower BOund maximization framework.
• We identify and implement two properties: the salient

space discriminability and the salient/common independence, that have not been successfully addressed by previous Contrastive VAE methods.
• We provide a fair comparison with other SOTA CA-VAE methods on 3 medical applications and a natural image experiment.

3. Contrastive Variational Autoencoders

Let $(X, Y) = \{(x_i, y_i)\}_{i=1}^N$ be a data-set of images x_i associated with labels $y_i \in \{0, 1\}$, 0 for background and 1 for target. Both background and target samples are assumed to be i.i.d. from two different and unknown distributions that depend on two latent variables: $c_i \in \mathbf{R}^{D_c}$ and $s_i \in \mathbf{R}^{D_s}$. Our objective is to have a generative model $x_i \sim p_\theta(x|y_i, c_i, s_i)$ so that: 1- the **common** latent vectors $C = \{c_i\}_{i=1}^N$ should capture the common generative factors of variation between the background and target distributions and fully encode the background samples and 2- the **salient** latent vectors $S = \{s_i\}_{i=1}^N$ should capture the distinct generative factors of variation of the target set (*i.e.,* patterns that are only present in the target dataset and not in the background dataset).

Similar to previous works(Abid & Zou, 2019; Weinberger et al., 2022; Zou et al., 2022), we assume the generative process: $p_\theta(x, y, c, s) = p_\theta(x|c, s, y)p_\theta(c)p_\theta(s|y)p(y)$. Since $p_\theta(c, s|x, y)$ is hard to compute in practice, we approximate it using an auxiliary parametric distribution $q_\phi(c, s|x, y)$ and directly derive the Evidence Lower Bound of $\log p(x, y)$.

Based on this generative latent variable model, one can derive the ELBO of the marginal log-likelihood $\log p(x, y)$,

$$-\log p_\theta(x, y) \leq \mathbf{E}_{c, s \sim q_{\phi_c, \phi_s}(c, s|x, y)} \log \frac{q_{\phi_c, \phi_s}(c, s|x, y)}{p_\theta(x, y, c, s)} \quad (1)$$

where we have introduced an auxiliary parametric distribution $q_\phi(c, s|x, y)$ to approximate $p_\theta(c, s|x, y)$.

From there, we can develop the lower bound into three terms, a conditional reconstruction term, a common space prior regularization, and a salient space prior regularization:

$$
\begin{aligned}
-\log p_\theta(x, y) \leq & -\underbrace{\mathbf{E}_{c,s \sim q_{\phi_c,\phi_s}(c,s|x,y)} \log p_\theta(x|y, c, s)}_{\textbf{Conditional Reconstruction}} \\
& + \underbrace{KL(q_{\phi_c}(c|x)||p_\theta(c))}_{\textbf{b) Common prior}} + \underbrace{KL(q_{\phi_s}(s|x, y)||p_\theta(s|y))}_{\textbf{c) Salient prior}}
\end{aligned}
\tag{2}
$$

Here, we assume the independence of the auxiliary distributions (*i.e.:* $q_{\phi_c,\phi_s}(c, s|x, y) = q_{\phi_c}(c|x)q_{\phi_s}(s|x, y)$) and prior distributions (*i.e.:* $p_\theta(c, s) = p_\theta(c)p_\theta(s)$). Both $p_\theta(x|y_i, c_i, s_i)$ (i.e., single decoder) and $q_{\phi_c}(c|x)q_{\phi_s}(s|x, y)$ (i.e., two encoders) are assumed to follow a Gaussian distribution parametrized by a neural network. To reinforce the independence assumption between c and s, we introduce a Mutual Information regularization term $KL(q(c, s)||q(c)q(s))$. Theoretically, this term is similar to the one in (Abid & Zou, 2019). This property is desirable in order to ensure that the information is well separated between the latent spaces. However, in (Abid & Zou, 2019), the Mutual Information estimation and minimization are done simultaneously [1]. In this paper, we argue that the estimation of the Mutual Information requires the introduction of an independent optimizer, see Sec. 3.5. To further reduce the overlap of target and common distributions on the salient space, we also introduce a salient classification loss defined as $\mathbf{E}_{s \sim q_{\phi_s}(s|x,y)} \log p(y|s)$.

By combining all these losses together, we obtain the final loss $\mathcal{L}$:

$$
\begin{aligned}
\mathcal{L} = & -\underbrace{\mathbf{E}_{c,s \sim q_{\phi_c,\phi_s}(c,s|x,y)} \log p_\theta(x|c, s, y)}_{\textbf{a) Conditional Reconstruction}} \\
& + \underbrace{KL(q(c, s)||q(c)q(s))}_{\textbf{e) Mutual Information}} - \underbrace{\mathbf{E}_{s \sim q_{\phi_s}(s|x,y)} \log p_\theta(y|s)}_{\textbf{d) Salient Classification}} \\
& + \underbrace{KL(q_{\phi_c}(c|x)||p_\theta(c))}_{\textbf{b) Common Prior}} + \underbrace{KL(q_{\phi_s}(s|x, y)||p_\theta(s|y))}_{\textbf{c) Salient Prior}}
\end{aligned}
\tag{3}
$$

3.1. Conditional reconstruction

The reconstruction loss term is given by $-\mathbf{E}_{c,s \sim q_{\phi_c,\phi_s}(c,s|x,y)} \log p_\theta(x|c, s, y)$. Given an image x (and a label y), a common and a salient latent vector can be drawn from q_{ϕ_c,ϕ_s} with the help of the reparameterization trick.

We assume that $p(x|c, s, y) \sim \mathcal{N}(d_\theta([c, ys + (1 - y)s'], I)$, *i.e:* $p_\theta(x|c, s, y)$ follows a Gaussian distribution parameterized by θ, centered on $\mu_{\hat{x}} = d_\theta([c, ys + (1 - y)s'])$ with identity covariance matrix, and d_θ is the decoder and $[., .]$ denotes a concatenation.

Therefore, by developing the reconstruction loss term, we obtain the mean squared error between the input and the reconstruction: $\mathcal{L}_{\text{rec}} = \sum_{i=1}^N ||x - d_\theta([c, ys + (1 - y)s'])||_2^2$. Importantly, for background samples, we set the salient latent vectors to $\mathbf{s'} = 0$. This choice enables isolating the background factors of variability in the common space only.

3.2. Common prior

By assuming $p(c) \sim \mathcal{N}(0, I)$ and $q_{\phi_c}(c|x) \sim \mathcal{N}(\mu_\phi(x), \sigma_\phi(x, y))$, the KL loss has a closed form solution, as in standard VAEs. Here, both $\mu_\phi(x)$ and $\sigma_\phi(x, y)$ are the outputs of the encoder e_{ϕ_c}.

3.3. Salient prior

To compute this regularization, we first need to develop $p_\theta(s) = \sum_y p(y)p_\theta(s|y)$, where we assume that $p(y)$ follows a Bernoulli distribution with probability equal to 0.5. Thus, the salient prior reduces to a formula that only depends on $p_\theta(s|y)$, which is conditioned by the knowledge of the label (0: background, 1: target). This allows us to distinguish between the salient priors of background samples ($p(s|y = 0)$) and target samples ($p(s|y = 1)$).

Similar to other CA-VAE methods, we assume that $p(s|y = 1) \sim \mathcal{N}(0, I)$ and , as in (Zou et al., 2022), that $p(s|x, y = 0) \sim \mathcal{N}(s', \sqrt{\sigma_p}I)$, with $s' = 0$ and $\sqrt{\sigma_p} < 1$, namely a Gaussian distribution centered on an informationless reference s' with a small constant variance σ_p. We preferred it to a Delta function $\delta(s = s')$ (as in (Weinberger et al., 2022)) because it eases the computation of the KL divergence (i.e., closed form) and it also means that we tolerate a small salient variation in the background (healthy) samples. In real applications, in particular medical ones, diagnosis labels can be noisy, and mild pathological patterns may exist in some healthy control subjects. Using such a prior, we tolerate these possible (erroneous) sources of variation. Furthermore, one could also extend the proposed method to a continuous y, for instance, between 0 and 1, describing the severity of the disease. Indeed, practitioners could define a function $\sigma_p(y)$ that would map the severity score y to a salient prior standard deviation (e.g., $\sigma_p(y) = y$). In this way, we could extend our framework to the case where pathological variations would follow a continuum from no (or mild) to severe patterns.

---

[1] In (Abid & Zou, 2019), Algorithm 1 suggests that the Mutual Information estimation and minimization depend on two distinct parameters update. However, in practice, in their code, a single optimizer is used. This is also confirmed in Sec. 3, where authors write: "discriminator is trained simultaneously with the encoder and decoder neural networks".

3.4. Salient classification

In the salient prior regularization, as in previous works, we encourage background and target salient factors to match two different Gaussian distributions, both centered in 0 (we assume $s' = 0$) but with different covariance. However, we argue that target salient factors should be further encouraged to differ from the background ones in order to reduce the overlap of target and common distributions on the salient space and enhance the expressivity of the salient space.

To encourage target and background salient factors to be generated from different distributions, we propose to minimize a Binary Cross Entropy loss to distinguish the target from background samples in the salient space. Assuming that $p(y|s)$ follows a Bernoulli distribution parameterized by $f_\xi(s)$, a 2-layers classification Neural Network, we obtain a Binary Cross Entropy (BCE) loss between true labels y and predicted labels $\hat{y} = f_\xi(s)$.

3.5. Mutual Information

To promote independence between c and s, we minimize their mutual information, defined as the KL divergence between the joint distribution $q(c, s)$ and the product of their marginals $q(c)q(s)$.

However, computing this quantity is not trivial, and it requires a few tricks in order to correctly estimate and minimize it. As in (Abid & Zou, 2019), it is possible to take inspiration from FactorVAE (Kim & Mnih, 2019), which proposes to estimate the density-ratio between a joint distribution and the product of the marginals. In our case, we seek to enforce the independence between two sets of latent variables rather than between each latent variable of a set. The density-ratio trick (Nguyen et al., 2010; Sugiyama et al., 2012) allows us to estimate the quantity inside the $\log$ in Eq.4. First, we sample from $q(c, s)$ by randomly choosing a batch of images (x_i, y_i) and drawing their latent factors $[c_i, s_i]$ from the encoders e_{ϕ_c} and e_{ϕ_s}. Then, we sample from $q(c)q(s)$ by using the same batch of images where we shuffle the latent codes among images (e.g., $[c_1, s_2], [c_2, s_3]$, etc.). Once we obtained samples from both distributions, we trained an **independent** classifier $D_\lambda([c, s])$ to discriminate the samples drawn from the two distributions by minimizing a BCE loss. The classifier is then used to approximate the ratio in the KL divergence, and we can train the encoders e_{ϕ_c} and e_{ϕ_s} to minimize the resulting loss:

$$\mathcal{L}_{\text{MI}} = \mathbb{E}_{q(c,s)} \log \left(\frac{q(c, s)}{q(c)q(s)} \right)$$
$$\approx \sum_i \text{ReLU} \left(\log \left(\frac{D_\lambda([c_i, s_i])}{1 - D_\lambda([c_i, s_i])} \right) \right) \quad (4)$$

where the ReLU function forces the estimate of the KL divergence to be positive, thus avoiding to back-propagate

wrong estimates of the density ratio due to the simultaneous training of $D_\lambda([c, s])$. In (Abid & Zou, 2019), while Alg.1 of the original paper describes two distinct gradient updates, it is written that "This discriminator is trained simultaneously with the encoder and decoder neural networks". In practice, a single optimizer is used in their training code. In our work, we use an independent optimizer for D_λ, in order to ensure that the density ratio is well estimated. Furthermore, we freeze D_λ's parameters when minimizing the Mutual Information estimate. The pseudo-code is available in Alg. 1, and a visual explanation is shown in Fig.3.
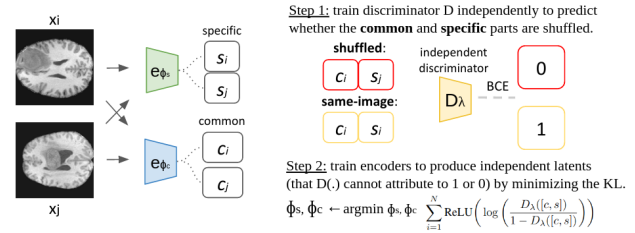

Figure 3. Illustration of Mutual Information loss between the common and the salient space. Given two images x_a and x_b, 4 sets of latents are computed: c_a and s_a latents of the image a, c_b and s_b latents of the image b. A non-linear MLP is independently trained with a binary cross-entropy loss to classify shuffled concatenations (i.e., from different images) with the label 0 and concatenations of latents coming from the same image with label 1. Then, during training, encoders should not to be able to identify whether a concatenation of latents belong to class 0 (shuffled common and salient spaces) or class 1 (common and salient spaces coming from the same image). We encourage that by minimizing $D_{KL}(p_{\phi_s, \phi_c}(c, s) || p_{\phi_c}(c) \times p_{\phi_s}(s))$.

Algorithm 1 Minimizing the Mutual Information between common and salient spaces, given a batch of size B.

1: **Input:** $X \in \mathbf{R}^{B \times (C \times W \times H)}$
2: **for** t in epochs : **do**
3:     Discriminator training :
4:     Sample $z = [c, s]$ from q_{ϕ_c, ϕ_s}.
5:     Sample $\bar{z} = [c, \bar{s}]$ from $q_{\phi_c} \times q_{\phi_s}$ by shuffling s along the batch dimension.
6:     Compute $\mathcal{L}_{BCE} = -\log(D(z)) - \log(1 - D(\bar{z}))$
7:     Freeze ϕ_c and ϕ_s. Update D parameters only.
8:     Encoders training :
9:     Sample $z = [e_{\phi_c}(x), e_{\phi_s}(x)]$ from q_{ϕ_c, ϕ_s}.
10:     Compute $\mathcal{L}_{MI} = \sum_{i=1}^{B} \text{ReLU}\left(\log \frac{D(z_i)}{1 - D(z_i)} \right)$
11:     Freeze D parameters. Update ϕ_c and ϕ_s.
12: **end for**

---

4. Experiments

4.1. Evaluation details

Here, we evaluate the ability of SepVAE to separate common from target-specific patterns on three medical and one natural (CelebA) imaging datasets. We compare it with the only SOTA CA-VAE methods whose code is available: MM-cVAE (Weinberger et al., 2022) and ConVAE [2] (Abid & Zou, 2019).

For quantitative evaluation, we use the fact that the information about attributes, clinical variables, or subtypes (e.g. glasses/hats in CelebA) should be present either in the common or in the salient space. Once the encoders/decoder are trained, we evaluate the quality of the representations in two steps. First, we train a Logistic (resp. Linear) Regression on the estimated salient and common factors of the training set to predict the attribute presence (resp. attribute value). Then, we evaluate the classification/regression model on the salient and common factors estimated from a test set. By evaluating the performance of the model, we can understand whether the information about the attributes/variables/subtype has been put in the common or salient latent space by the method. Furthermore, we report the background (BG) vs target (TG) classification accuracy. To do so, a 2 layers MLPs is independently trained, except for SepVAE, where salient space predictions are directly estimated by the classifier.

In all Tables, for categorical variables, we compute (Balanced) Accuracy scores (=(B-)ACC), or Area-under Curve scores (=AUC) if the target is binary. For continuous variables, we use Mean Average Error (=MAE). Best results are highlighted in bold, second best results are underlined. For CelebA and Pneumonia experiments, mean, and standard deviations are computed on the results of 5 different runs in order to account for model initializations. For neuropsychiatric experiments, mean and standard deviations are computed using a 5-fold cross-validation evaluation scheme.

Qualitatively, the model can be evaluated by looking at the full image reconstruction (common+salient factors) and by fixing the salient factors to s' for target images. Comparing full reconstructions with common-only reconstructions allows the user to interpret the patterns encoded in the salient factors s (see Fig.1 and Fig.5).

4.2. CelebA - glasses vs hat identification

To compare with (Weinberger et al., 2022), we evaluated our performances on the CelebA with attributes dataset. It contains two sets, target and background, from a subset of CelebA (Liu et al., 2015), one with images of celebrities wearing glasses or hats (target) and the other with images

---

[2]ConVAE implemented with correct Mutual Information minimization, *i.e.:* with independently trained discriminator.

Figure 4. CelebA accessories dataset. We used a train set of 20000 images (10000 no accessories, 5000 glasses, 5000 hats) and an independent test set of 4000 images (2000 no accessories, 1000 glasses, 1000 hats) and ran the experiment 5 times to account for initialization uncertainty. Images were centered on the face and then resized to 64×64, pixels were normalized between 0 and 1.

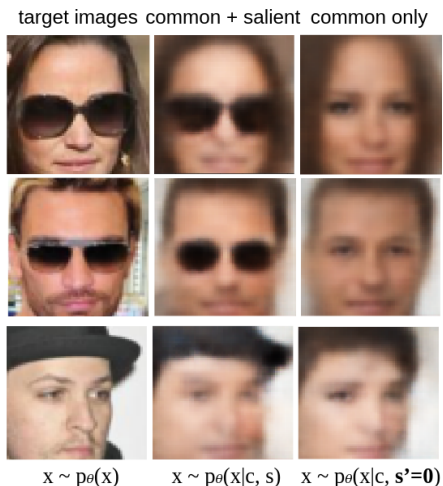$x \sim p_\theta(x)$ $x \sim p_\theta(x|c, s)$ $x \sim p_\theta(x|c, \mathbf{s'=0})$

Figure 5. SepVAE qualitative example on the CelebA with accessories dataset (BG = no accessories, TG = hats and glasses). (Middle, common+salient): Full reconstructions using the estimated common and salient factors. (Right, common only): Reconstruction using only the estimated common factors fixing the salient to s'. The salient latent variables capture the accessories (hats and glasses), which are target-specific patterns. The common latents capture the common attributes (e.g., identity, skin color).

Table 1. CA-VAE methods performance on CelebA with accessories dataset. Accessories (glasses/hat) information should only be present in the salient space, not in the common.

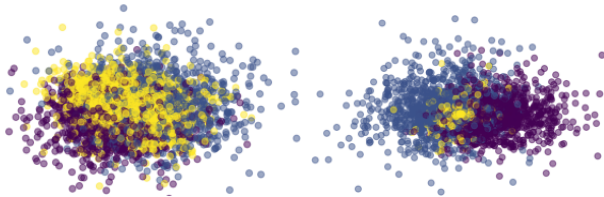|	Glss/Hats Acc SALIENT ↑	Glss/Hats Acc COMMON ↓	Bg vs Tg AUC SALIENT ↑	Bg vs Tg AUC COMMON ↓
ConVAE	82.32±1.17	75.01±2.526	82.46±0.586	78.39±0.41
MM-cVAE	85.17±0.60	73.938±1.66	88.536±0.39	78.036±0.35
SepVAE	**87.62±0.75**	**72.16±2.02**	**93.15±1.65**	**77.604±0.20**

Figure 6. PCA projections of MM-c-VAE (left) and SepVAE (right) salient space on CelebA TEST set. Yellow: no accessories. Dark Blue: glasses. Purple: hats. We can clearly observe that our method maximizes the target variance while reducing the background variance. We attribute this different behaviour to our salient classification loss, which reduces the overlap between background and target salient distributions.

of celebrities not wearing these accessories (background). The discriminative information allowing the classification of glasses vs. hats should only be present in the salient latent space. We demonstrate that we successfully encode these attributes in the salient space with quantitative results in Tab. 1, and with reconstruction results in Fig. 5. Furthermore, in Fig. 6, we show that we effectively minimize the background dataset variance in the salient space compared to MM-cVAE[3].

4.3. Identify pneumonia subgroups

We gathered 1342 healthy X-Ray radiographies (*background* dataset), 2684 radiographies of pneumonia radiographies (*target* dataset) from (Kermany et al., 2018). Two different sub-types of pneumonia constitute this set, viral (1342 samples) and bacterial (1342 samples), see Fig.7. Radiographies were selected from a cohort of pediatric patients aged between one and five years old from Guangzhou Women and Children's Medical Center, Guangzhou. TRAIN set images were graded by 2 radiologists experts and the independent TEST set was graded by a third expert to account for label uncertainty. In Tab. 2, we demonstrate that our method is able to produce a salient space that captures the pathological variability as it allows distinguishing the two subtypes: viral and bacterial pneumonia.

---

[3]Our evaluation process is different from (Weinberger et al., 2022) as their TEST set has been used during the model training. Indeed, the TRAIN / TEST split used for training Logistic Regression is performed after the model fitting on the set TRAIN+TEST set. Besides, we were not able to reproduce their results.

Figure 7. Illustration of the pneumonia dataset. Target images are pneumonia images composed of viral and bacterial pneumonia. Background images are healthy X-Ray images. Original dataset image description from (Kermany et al., 2018). The dataset is available at https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia.

Table 2. CA-VAE methods performance on the Healthy vs Pneumonia X-Ray dataset. Accuracy scores are obtained with linear probes fitted on common c or salient s latent vectors of the images of the target dataset. Pneumonia subtypes information should only be present in the salient space. The lower part shows an ablation study of regularization losses.

	Subgrp Acc SALIENT ↑	Subgrp Acc COMMON ↓	Bg vs Tg Acc SALIENT ↑	Bg vs Tg Acc COMMON ↓
ConVAE	82.30±1.53	73.58±1.84	67.80±5.93	58.05±7.17
MM-cVAE	82.86±1.87	74.35±3.19	70.44±2.69	59.94±5.88
SepVAE	**84.78±0.42**	**70.92±1.39**	**78.13±3.03**	57.52±4.14
SepVAE no **MI**	84.10±0.48	71.792±2.94	75.186±5.69	60.35±4.73
SepVAE no **CLSF**	84.71±1.19	73.58±2.19	71.91±4.65	**55.79±5.41**
SepVAE no **REG**	83.98±0.85	72.61±2.05	73.03±2.97	61.43±2.25

Ablation study In the lower part of Tab. 2, we propose to disable different components of the model to show that the full model SepVAE is always better on average. no **MI** means that we disabled the Mutual Information minimization loss (no Mutual Information Minimization). no **CLSF** means that we disabled the classification loss on the salient space (no Salient Classification). no **REG** means that we disabled the regularization loss that forces the background samples to align with an informationless vector **s'** = 0 (no Salient Prior).

Table 3. CA-VAE methods performance on the prediction of schizophrenia-specific variables (SANS, SAPS, Diag) and common variables (Age, Sex, Site) using only salient factors reconstructed by test images of the target (MD) dataset.

	Age MAE ↑	Sex B-Acc ↓	Site B-Acc ↓	SANS MAE ↓	SAPS MAE ↓	Diag AUC ↑
ConVAE	7.46±0.18	72.72±1.32	54.46±2.46	**3.95±0.28**	2.76±0.18	58.53±4.87
MM-cVAE	7.10±0.34	72.15±2.47	56.69±9.84	4.52±0.33	3.16±0.05	70.94±4.08
SepVAE	**7.98±0.25**	72.61±2.19	**44.10±5.78**	4.14±0.39	**2.60±0.27**	**79.15±3.39**

Table 4. CA-VAE methods performance on the prediction of autism-specific variables (ADOS (Akshoomoff et al., 2006), ADI-s, Diag) and common variables (Age, Sex, Site) using only salient factors reconstructed by test images of the target (MD) dataset.

	Age MAE ↑	Sex B-Acc ↓	Site B-Acc ↓	ADOS MAE ↓	ADI-s MAE ↓	Diag AUC ↑
ConVAE	3.97±0.19	66.67±1.12	40.97±2.06	10.1±1.27	5.14±0.17	54.93±2.04
MM-cVAE	3.74±0.12	64.07±2.58	40.93±2.66	10.5±2.47	5.09±0.16	54.88±2.76
SepVAE	**4.38±0.09**	59.61±1.78	33.58±1.86	**8.55±1.68**	4.91±0.17	**59.73±1.78**

4.4. Parsing neuro-anatomical variability in psychiatric diseases

The task of identifying consistent correlations between neuro-anatomical biomarkers and observed symptoms in psychiatric diseases is important for developing more precise treatment options. Separating the different latent mechanisms that drive neuro-anatomical variability in psychiatric disorders is a challenging task. Contrastive Analysis (CA) methods such as ours have the potential to identify and separate healthy from pathological neuro-anatomical patterns in structural MRIs. This ability could be a key component to push forward the understanding of the mechanisms that underlie the development of psychiatric diseases.

Given a background population of Healthy Controls (HC) and a target population suffering from a Mental Disorder (MD), the objective is to capture the pathological factors of variability in the salient space, such as psychiatric and cognitive clinical scores, while isolating the patterns related to demographic variables, such as age and sex, or acquisition sites to the common space. For each experiment, we gather T1w anatomical VBM (Ashburner & Friston, 2000) preprocessed images resized to 128x128x128 of HC and MD subjects. We divide them into 5 TRAIN, VAL splits (0.75, 0.25) and evaluate in a cross-validation scheme the performance of SepVAE and the other SOTA CA-VAE methods. Please note that this is a challenging problem, especially due to the high dimensionality of the input and the scarcity of the data. Notably, the measures of psychiatric and cognitive clinical scores are only available for some patients, making it scarce and precious information.

4.4.1. SCHIZOPHRENIA:

We merged images of schizophrenic patients (TG) and healthy controls (BG) from the datasets SCHIZCONNECT-VIP (Wang et al., 2016) and BSNIP (Tamminga et al., 2014). Results in Tab. 3 show that the salient factors estimated using our method better predict schizophrenia-specific variables of interest: SAPS (Scale of Positive Symptoms), SANS (Scale of Negative Symptoms), and diagnosis. On the other hand, salient features are shown to be poorly predictive of demographic variables: age, sex, and acquisition site. It paves the way toward a better understanding of schizophrenia disorder by capturing neuro-anatomical patterns that are predictive of the psychiatric scales while not being biased by confound variables.

4.4.2. AUTISM:

Second, we combine patients with autism from ABIDE1 and ABIDE2 (Heinsfeld et al., 2017) (TG) with healthy controls (BG). In Tab. 4, SepVAE's salient latents better predict the diagnosis and the clinical variables, such as ADOS (Autism Diagnosis Observation Schedule) and ADI Social (Autism Diagnosis Interview Social) which quantifies the social interaction abilities. On the other hand, salient latents poorly infer irrelevant demographic variables (age, sex, and acquisition site), which is a desirable feature for the development of unbiased diagnosis tools.

5. Conclusions and Perspectives

In this paper, we developed a novel CA-VAE method entitled SepVAE. Building onto Contrastive Analysis methods, we first criticize previously proposed regularizations about (1) the matching of target and background distributions in the common space and (2) the overlapping of target and background priors in the salient space. These regularizations may fail to prevent information leakage between common and salient spaces, especially when datasets are biased. We thus propose two alternative solutions: salient discrimination between target and background samples, and mutual information minimization between common and salient spaces. We integrate these losses along with the maximization of the ELBO of the joint log-likelihood. We demonstrate superior performances on radiological and two neuro-psychiatric applications, where we successfully separate the pathological information of interest (diagnosis, pathological scores) from the "nuisance" common variations (e.g., age, site). The development of methods like ours seems very promising and offers a large spectrum of perspectives. For example, it could be further extended to multiple target datasets (*e.g.,* healthy population Vs several pathologies, to obtain a continuum healthy - mild - severe pathology) and to other models, such as GANs, for improved generation quality. Eventually, to be entirely trustworthy, the model must be identifiable, namely, we need to know the conditions that allow us to learn the correct joint distribution over observed and latent variables. We plan to follow (Khemakhem et al., 2020; von Kügelgen et al., 2021) to obtain theoretic guarantees of identifiability of our model.

References

Abid, A. and Zou, J. Contrastive Variational Autoencoder Enhances Salient Features, February 2019. arXiv:1902.04601 [cs, stat].

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, May 2018. ISSN 2041-1723.

Ainsworth, S. K., Foti, N. J., Lee, A. K. C., and Fox, E. B. oi-VAE: Output Interpretable VAEs for Nonlinear Group Factor Analysis. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 119–128. PMLR, July 2018. ISSN: 2640-3498.

Akshoomoff, N., Corsello, C., and Schmidt, H. The Role of the Autism Diagnostic Observation Schedule in the Assessment of Autism Spectrum Disorders in School and Community Settings. *The California school psychologist: CASP*, 11:7–19, 2006. ISSN 1087-3414.

Antelmi, L., Ayache, N., Robert, P., and Lorenzi, M. Sparse multi-channel variational autoencoder for the joint analysis of heterogeneous data. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 302–311. PMLR, 09–15 Jun 2019.

Ashburner, J. and Friston, K. J. Voxel-based morphometry–the methods. *NeuroImage*, 11(6 Pt 1):805–821, June 2000. ISSN 1053-8119.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-VAE. *arXiv:1804.03599 [cs, stat]*, April 2018.

Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating Sources of Disentanglement in Variational Autoencoders. *arXiv:1802.04942 [cs, stat]*, April 2019.

Choudhuri, A., Makkuva, A. V., Rana, R., Oh, S., Chowdhary, G., and Schwing, A. Towards Principled Objectives for Contrastive Disentanglement. December 2019.

Ge, R. and Zou, J. Rich component analysis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pp. 1502–1510, New York, NY, USA, June 2016. JMLR.org.

Haber, A. L., Biton, M., Rogel, N., Herbst, R. H., Shekhar, K., Smillie, C., Burgin, G., Delorey, T. M., Howitt, M. R., Katz, Y., Tirosh, I., Beyaz, S., Dionne, D., Zhang, M., Raychowdhury, R., Garrett, W. S., Rozenblatt-Rosen, O., Shi, H. N., Yilmaz, O., Xavier, R. J., and Regev, A. A single-cell survey of the small intestinal epithelium. *Nature*, 551(7680):333–339, November 2017. ISSN 1476-4687. Number: 7680 Publisher: Nature Publishing Group.

Heinsfeld, A. S., Franco, A. R., Craddock, R. C., Buchweitz, A., and Meneguzzi, F. Identification of autism spectrum disorder using deep learning and the ABIDE dataset. *NeuroImage : Clinical*, 17:16–23, August 2017. ISSN 2213-1582.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*, 2017.

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., Holtzman, D. M., Jagust, W., Jessen, F., Karlawish, J., Liu, E., Molinuevo, J. L., Montine, T., Phelps, C., Rankin, K. P., Rowe, C. C., Scheltens, P., Siemers, E., Snyder, H. M., Sperling, R., and Contributors. NIA-AA Research Framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 14(4): 535–562, April 2018. ISSN 1552-5279.

Jones, A., Townes, F. W., Li, D., and Engelhardt, B. E. Contrastive latent variable modeling with application to case-control sequencing experiments, February 2021.

Joy, T., Schmon, S. M., Torr, P. H. S., Siddharth, N., and Rainforth, T. Capturing Label Characteristics in VAEs. *ICLR 2021*, June 2021.

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M. Y. L., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., Shi, A., Zhang, R., Zheng, L., Hou, R., Shi, W., Fu, X., Duan, Y., Huu, V. A. N., Wen, C., Zhang, E. D., Zhang, C. L., Li, O., Wang, X., Singer, M. A., Sun, X., Xu, J., Tafreshi, A., Lewis, M. A., Xia, H., and Zhang, K. Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, 172(5):1122–1131.e9, February 2018. ISSN 0092-8674, 1097-4172. Publisher: Elsevier.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *AISTATS*, 2020.

Kim, H. and Mnih, A. Disentangling by Factorising, July 2019. NeurIps 2017.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. December 2013.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised Learning with Deep Generative Models.

Li, Y., Pan, Q., Wang, S., Peng, H., Yang, T., and Cambria, E. Disentangled Variational Auto-Encoder for Semi-supervised Learning. *arXiv:1709.05047 [cs]*, December 2018.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep Learning Face Attributes in the Wild, September 2015. arXiv:1411.7766 [cs].

Maaløe, L., Sønderby, C. K., Sønderby, S. K., and Winther, O. Auxiliary Deep Generative Models. In *Proceedings of The 33rd International Conference on Machine Learning*, pp. 1445–1453. PMLR, June 2016. ISSN: 1938-7228.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. Disentangling Disentanglement in Variational Autoencoders. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4402–4412. PMLR, May 2019. ISSN: 2640-3498.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., and et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, October 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, November 2010. ISSN 0018-9448, 1557-9654.

Ruiz, A., Martinez, O., Binefa, X., and Verbeek, J. Learning Disentangled Representations with Reference-Based Variational Autoencoders, January 2019. arXiv:1901.08534 [cs].

Severson, K. A., Ghosh, S., and Ng, K. Unsupervised Learning with Contrastive Latent Variable Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 4862–4869, July 2019. ISSN 2374-3468.

Shu, R., Zhao, S., and Kochenderfer, M. J. Rethinking style and content disentanglement in variational auto encoders. 2018. ICLR Workshop.

Sugiyama, M., Suzuki, T., and Kanamori, T. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64:1009–1044, 2012.

Tamminga, C. A., Pearlson, G., Keshavan, M., Sweeney, J., Clementz, B., and Thaker, G. Bipolar and Schizophrenia Network for Intermediate Phenotypes: Outcomes Across the Psychosis Continuum. *Schizophrenia Bulletin*, 40 (Suppl 2):S131–S137, March 2014. ISSN 0586-7614.

von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-Supervised Learning with Data Augmentations Provably Isolates Content from Style. In *NeurIPS*, 2021.

Wang, L., Alpert, K. I., Calhoun, V. D., Cobia, D. J., Keator, D. B., King, M. D., Kogan, A., Landis, D., Tallis, M., Turner, M. D., Potkin, S. G., Turner, J. A., and Ambite, J. L. SchizConnect: Mediating neuroimaging databases on schizophrenia and related disorders for large-scale integration. *NeuroImage*, 124(Pt B):1155–1167, January 2016. ISSN 1095-9572.

Weinberger, E., Beebe-Wang, N., and Lee, S.-I. Moment Matching Deep Contrastive Latent Variable Models. In *AISTATS*, 2022.

Zheng, G. X. Y., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., Bharadwaj, R., Wong, A., Ness, K. D., Beppu, L. W., Deeg, H. J., McFarland, C., Loeb, K. R., Valente, W. J., Ericson, N. G., Stevens, E. A., Radich, J. P., Mikkelsen, T. S., Hindson, B. J., and Bielas, J. H. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8(1):14049, January 2017. ISSN 2041-1723. Number: 1 Publisher: Nature Publishing Group.

Zheng, Z. and Sun, L. Disentangling Latent Space for VAE by Label Relevant/Irrelevant Dimensions. *arXiv:1812.09502 [cs]*, March 2019. arXiv: 1812.09502.

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*, volume 26. Curran Associates, Inc., 2013.

Zou, K., Faisan, S., Heitz, F., and Valette, S. Joint Disentanglement of Labels and Their Features with VAE. In *IEEE International Conference on Image Processing (ICIP)*, pp. 1341–1345, 2022.

Supplementary

A. Context on Variational Auto-Encoders

Variational Autoencoders (VAEs) are a type of generative model that can be used to learn a compact, continuous latent representation of a dataset. They are based on the idea of using an encoder network to map input data points x (*e.g:* an image) to a latent space z, and a decoder network to map points in the latent space back to the original data space.

Mathematically, given a dataset $X = x_i{}_{i=1}^N$ and a VAE model with encoder $q_\phi(z|x)$ and decoder $p_\theta(x|z)$, the VAE seeks ϕ, θ to maximize a lower bound of the input distribution likelihood:

$$\log p_\theta(x) \leq \mathbf{E}_{z \sim q_\phi(z|x)} \log p_\theta(x|z) - KL(q_\phi(z|x)||p_\theta(z))$$

where $p_\theta(x|z)$ is the likelihood of the input space, and $\mathrm{KL}(q_\phi(z|x)||p(z))$ is the Kullback-Leibler divergence between $q_\phi(z|x)$, the approximation of the posterior distribution, and $p(z)$ the prior over the latent space (often chosen to be a standard normal distribution).

The first term in the objective function, $\mathbf{E}_{z \sim q_\phi(z|x)} \log p_\theta(x|z)$, is the negative reconstruction error, which measures how well the decoder can reconstruct the input data from the latent representation. The second term, $\mathrm{KL}(q_\phi(z|x)||p(z))$, encourages the encoder distribution to be similar to the prior distribution, which helps to prevent overfitting and encourage the learned latent representation to be continuous and smooth.

B. Salient posterior sampling for background samples

In Sec. 3.3, we motivated the choice of a peaked Gaussian prior for salient background distribution with a user-defined σ_p. This way, the derivation of the Kullback-Leiber divergence is directly analytically tractable as in standard VAEs.
To simplify the optimization scheme, we could also set and freeze the standard deviations $\sigma_q^{y=0}$ of the salient space of the background samples. This way, it reduces the Kullback-Leiber divergence between $q_\phi(s|x, y=0)$ and $p_\theta(s|x, y=0)$ to a $\frac{1}{\sigma_p}$-weighted Mean Squared Error between $\mu_s(x|y=0)$ and s' : $\frac{||\mu_s^{x_i|y=0} - s'||_2^2}{\sigma_p}$. In our code, we make this choice as it simplifies the training scheme ($\sigma_q^{y=0}$ does not need to be estimated). In the case where there exists a continuum between healthy and diseased populations, $\sigma_q^{y=0}$ should be estimated.

Also, the choice of a frozen $\sigma_q^{y=0}$ allows controlling the radius of the classification boundary between background and target samples in the salient space. Indeed, the classifier is fed with samples from the target distributions ($q_{\phi_s(s|x, y=1)} \sim N(\mu_s(x), \sigma_s(x))$), and background distributions ($q_{\phi_s(s|x, y=0)} \sim N(\mu_s(x|y=0), \sigma_q)$). This implicitly avoids the overlap of both distributions with a margin proportional to σ_q. See Fig. 8 for a visual explanation.

C. Implementation Details

C.1. CelebA glasses and hat versus no accessories

We used a train set of 20000 images, (10000 no accessories, 5000 glasses, 5000 hats) and an independent test set of 4000 images (2000 no accessories, 1000 glasses, 1000 hats), and ran the experiment 5 times to account for initialization uncertainty. Images are of size 64×64, pixel were normalized between 0 and 1. For this experiment, we use a standard encoder architecture composed of 5 convolutions (channels 3, 32, 32, 64, 128, 256), kernel size 4, stride 2, and padding (1, 1, 1, 1, 1). Then, for each mean and standard deviations predicted (common and salient) we used two linear layers going from 256 to hidden size 32 to (common and salient) latent space size 16. The decoder was set in a symmetrical manner. We used the same architecture across all the concurrent works we evaluated. We used a common and latent space dimension of 16 each. The learning rate was set to 0.001 with an Adam optimizer. Oddly we found that re-instantiating it at each epoch led to better results (for concurrent works also), we think that it is because it forgets momentum internal states between the epochs. The models were trained during 250 epochs. To note, MM-cVAE used latent spaces of 16 (salient space) and 6 common space and a different architecture but we noticed that it led to artifacts in the reconstruction (see original contribution). Also, we did not succeed to reproduce their performances with their code, their model, and their latent spaces, even with the same experimental setup. We, therefore, used our model setting which led to better performances across each method with batch size equal to 512. We used $\beta_c = 0.5$ and $\beta_s = 0.5$, $\kappa = 2$, $\gamma = 1e - 10$, $\sigma_p = 0.025$. For MM-cVAE we used the same learning rate, $\beta_c = 0.5$ and $\beta_s = 0.5$, the background salient regularization weight 100, common regularization weight of
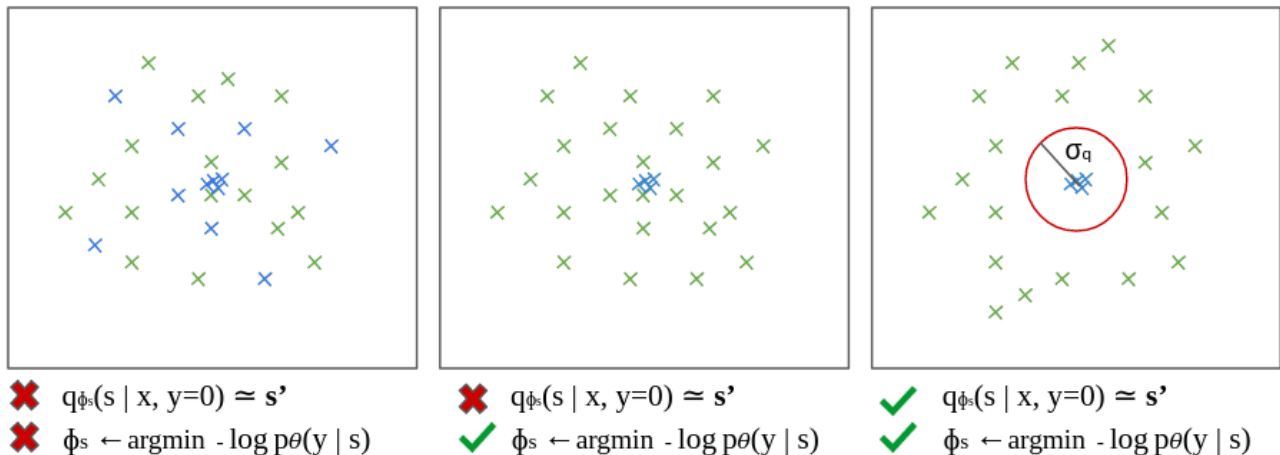
$\quad$ ✖ $\quad q_{\phi_s}(s \mid x, y=0) \simeq$ **s'** $\qquad$ ✖ $\quad q_{\phi_s}(s \mid x, y=0) \simeq$ **s'** $\qquad$ ✔ $\quad q_{\phi_s}(s \mid x, y=0) \simeq$ **s'**

$\quad$ ✖ $\quad \phi_s \leftarrow \text{argmin} - \log p_\theta(y \mid s)$ $\quad$ ✔ $\quad \phi_s \leftarrow \text{argmin} - \log p_\theta(y \mid s)$ $\quad$ ✔ $\quad \phi_s \leftarrow \text{argmin} - \log p_\theta(y \mid s)$

Figure 8. Illustration of the regularization loss within the salient space. As in MM-cVAE, the prior $q_{\phi_s(s \mid x, y=0)} \sim$ **s'** on the background samples (blue) forces their variance to be as small as possible. However, as the prior on target samples (green) follow a normal distribution, they may overlap with the background distribution. To avoid this case, our method trains a non-linear classifier to avoid the overlap of both distributions with a margin proportional to σ_q.

1000.

C.2. Pneumonia

Train set images were graded by 2 radiologists experts and the independent test set was graded by a third expert, the experiment was run 5 times to account for initialization uncertainty. Images are of size 64×64, pixel were normalized between 0 and 1. For this experiment, we use a standard encoder architecture composed of 4 convolutions (channels 3, 32, 32, 32, 256), kernel size 4, and padding (1, 1, 1, 0). Then, for each mean and standard deviations predicted (common and salient) we used two linear layers going from 256 to hidden size 256 to (common and salient) latent space size 128. The decoder was set in a symmetrical manner. We used the same architecture across all the concurrent works we evaluated. We used a common and latent space dimension of 128 each. The learning rate was set to 0.001 with an Adam optimizer. Oddly we found that re-instantiating it at each epoch led to better results (for concurrent works also), we think that it is because it forgets momentum internal states between the epochs. The models were trained during 100 epochs with batch size equal to 512. We used $\beta_c = 0.5$ and $\beta_s = 0.1$, $\kappa = 2$, $\gamma = 5e - 10$, $\sigma_p = 0.05$. For MM-cVAE, we used the same learning rate, $\beta_c = 0.5$ and $\beta_s = 0.1$, the background salient regularization weight 100, common regularization weight of 1000.

C.3. Neuro-psychiatric experiments

Images are of size $128 \times 128 \times 128$ with voxels normalized on a Gaussian distribution per image. Experiments were run 3 times with a different train/val/test split to account for initialization and data uncertainty. For this experiment, we use a standard encoder architecture composed of 3 3D-convolutions (channels 1, 32, 64, 128), kernel size 3, stride 2, and padding 1 followed by batch normalization layers. Then, for each mean and standard deviations predicted (common and salient), we used two linear layers going from 65536 to hidden size 256 to (common and salient) latent space size 128. The decoder was set symmetrically, except that it has four transposed convolutions (channels 128, 64, 32, 16, 1), kernel size 3, stride 2, and padding 1 followed by batch normalization layers. We used the same architecture across all the concurrent works we evaluated. We used a common and latent space dimension of 128 each. The models were trained during 51 epochs with a batch size equal to 32 with an Adam optimizer. For the Schizophrenia experiment, for Sep VAE, we used a learning rate of 0.00005, $\beta_c = 1$ and $\beta_s = 0.1$, $\kappa = 10$, $\gamma = 1e - 8$, $\alpha = \frac{1}{0.01}$. For MM-cVAE we used the same learning rate, $\beta_c = 1$ and $\beta_s = 0.1$, the background salient regularization weight 100, common regularization weight of 1000. For the Autism disorder experiment, we used a learning rate of 0.00002, $\beta_c = 1$ and $\beta_s = 0.1$, $\kappa = 10$, $\gamma = 1e - 8$, $\sigma_p = 0.01$. For MM-cVAE we used the same learning rate, $\beta_c = 1$ and $\beta_s = 0.1$, the background salient regularization weight 100, common regularization weight of 1000.