STYLE-CONTENT DISENTANGLEMENT UNDER CONDITIONAL SHIFT

Dan Andrei Iliescu & Damon J. Wischik [*]
Department of Computer Science and Technology
University of Cambriddge, UK
{dai24,damon.wischik}@cam.ac.uk

ABSTRACT

We propose a novel representation learning method called the Context-Aware Variational Autoencoder (CxVAE). Our model can perform style-content disentanglement on datasets with conditional shift. Conditional shift occurs when the distribution of a target variable $\mathbf{y}$ conditional on the input observation $\mathbf{x}$ — $p(\mathbf{y}|\mathbf{x})$ — changes across data environments (i.e. $p_i(\mathbf{y}|\mathbf{x}) \neq p_j(\mathbf{y}|\mathbf{x})$, where i, j are two different environments). We introduce two novel style-content disentanglement datasets to show empirically that existing methods fail to disentangle under conditional shift. We propose CxVAE, a model that overcomes this limitation by enforcing independence across the content variables inferred from each environment. Our model presents two innovations: a context-aware encoder and a content adversarial loss. We use a specially designed experiment to show empirically that these design choices directly cause an improvement of our model's performance on datasets with conditional shift.

1 INTRODUCTION

Style-content disentanglement is the machine learning task of decomposing observations into group-level and instance-level variation. Consider a dataset of D-dimensional observations organised into N groups $\{\mathbf{x}_{n,k} \ ; \ \mathbf{x}_{n,k} \in \mathbb{R}^D\}_{k=1}^{K_n}$ of size K_n where $n \in \{1 : N\}$. For simplicity, we use the notation $\mathbf{x}_{n,1:K_n}$ to denote the group. These observations could refer to many kinds of data: paintings grouped by artist, medical records grouped by patient, or economic data grouped by country. The "content" of an observation is defined as the set of attributes that vary within groups, while the "style" is defined as the attributes of the observation that vary across groups. To disentangle style and content, a representation network $r : \mathbb{R}^{D\times *} \to \mathbb{R}^{S+C\times *}$ is trained to encode one group of observations $\mathbf{x}_{n,1:K_n}$ into one style latent code $\mathbf{c}_n \in \mathbb{R}^S$ and a group of content latent codes $\mathbf{s}_{n,1:K_n} \in \mathbb{R}^{C\times *}$, one for each observation. The goal is for the style codes $\mathbf{s}$ to capture only the variation across groups and the content codes $\mathbf{c}$ to capture only the variation within groups.

Style-content disentanglement is crucial for real-world situations that involve groups of high-dimensional data. The low-dimensional representation of the content $\mathbf{c}$ can replace the underlying observation $\mathbf{x}$ as a more easily "digestible" input for a wide variety of ML tasks, such as classification, regression, clustering, and visualisation. Example applications benefitting from style-content disentanglement include computer vision (Tenenbaum & Freeman, 2000), data anonymisation (Louizos et al., 2016), and clinical risk modelling (Pavlou et al., 2015), among others.

Content representations are particularly useful when the data suffers conditional shift between groups. Conditional shift, also known as $Y|X$-shift, is a phenomenon that occurs when the distribution of some target variable $\mathbf{y}$ conditioned on the observation $\mathbf{x}$ changes from one group to another: $p(\mathbf{y}_n|\mathbf{x}_n) \neq p(\mathbf{y}_m|\mathbf{x}_m)$, where $n \neq m$ (Zhang et al., 2013b). In this situation, the parameter resulting from regressing the variable $\mathbf{y}$ on the observation $\mathbf{x}$ will generalise poorly from group n to group m. For example, a resting heart rate of 140 beats-per-minute is normal for a 2 year-old but dangerous for a 30 year-old. In this example, the heart rate is the observation $\mathbf{x}$, the medical severity is the variable $\mathbf{y}$, and the patient's age-bracket constitutes the group. Using a content representation

[*] Website: https://dan-andrei-iliescu.github.io/

$\mathbf{c}$ as the input to the classifier can potentially overcome this challenge since the content will be independent of the age-bracket by construction.

Even though conditional shift is widespread in real-world datasets, it is virtually absent from the datasets commonly used in the literature to evaluate style-content disentanglement (Liu et al., 2023). All the attributes of interest in the most popular datasets such as Shapes3D (Kim & Mnih, 2018), SmallNORB (LeCun et al., 2004), dSprites (Higgins et al., 2017), Cars3D (Reed et al., 2015), MPI3D (Gondal et al., 2019) can be identified easily through standard regression without the need to transform the raw observation to a style-neutral content representation. For example, the joint distribution over shapes, colours, and positions in dSprites (Higgins et al., 2017) stays constant regardless of which feature we choose to group by. The absence of conditional shift from these datasets hides the true generalisation performance of the state-of-the-art style-content disentanglement methods.

In our work, we show that state-of-the-art models fail to learn disentangled representations of style and content under conditional shift. We choose a representative selection of models: GVAE (Bouchacourt et al., 2018; Hosoya, 2019), AdaGVAE (Locatello et al., 2020), and COCO-FUNIT (Saito et al., 2020). We propose two new datasets as initial benchmarks for studying conditional shift in generative models: 1) A dataset of teapots viewed under different lighting conditions. The goal is to disentangle the colour of the teapot from the colour of the light (Figure 1). We call this dataset `Shift3DIdent`, since it is built upon the popular 3DIdent dataset (Zimmermann et al., 2021). 2) A synthetic dataset of standardised student test scores grouped by school. The goal is to disentangle student aptitude from the socio-economic factors associated with the school (Figure 2a). We call this dataset `TestScores`.

We introduce a model called the Context-Aware Variational Autoencoder (CxVAE), which is able to perform style-content disentanglement on datasets with severe conditional shift. This model exhibits two key differences from the Group VAE (GVAE) (Bouchacourt et al., 2018; Hosoya, 2019): 1) An encoder architecture that infers the content variable $\mathbf{c}$ conditionally on the style variable $\mathbf{s}$. 2) An adversarial loss that constrains the distribution of content variables in a group $\mathbf{c}_{1:K}$ to be mutually independent of one another. We show empirically that our method outperforms the set of existing models with respect to multiple disentanglement metrics.

We investigate the cause of the improvement in disentanglement performance brought by our model. Focusing on the `TestScores` dataset, we vary the strength of the conditional shift effect and re-train GVAE and CxVAE on each setting. After re-evaluation, we observe that the performance of the models is evenly matched when the conditional shift effect is negligible. However, as the strength of the conditional shift effect increases, the performance of GVAE decreases significantly. This shows that CxVAE addresses the problem of conditional shift directly.

Our contribution in this work is threefold: 1) We introduce two new datasets — `TestScores` and `Shift3DIdent` — for evaluating style-content disentanglement under conditional shift. These are a first step towards a unified benchmark for conditional shift. 2) We propose, implement, and evaluate a new method for performing style-content disentanglement,the Context-Aware Variational Autoencoder. Full details of the implementation, datasets, metrics, and experiments can be found on the GitHub repository [1]. 3) We show that the design of our model directly addresses the problem of conditional shift.

2 Related Work

Style-Content Disentanglement. This problem is known as style-content disentanglement (Tenenbaum & Freeman, 2000), content-transformation disentanglement (Hosoya, 2019), and disentanglement with group supervision (Shu et al., 2020), to name a few. Recent work (Shu et al., 2020; Locatello et al., 2020) has contextualised group disentanglement as a subproblem of weakly-supervised disentanglement, where disentangled representations are learned with the help of non-datapoint supervision (e.g., grouping, ranking, restricted labelling). Early work in this area focused on separating between visual concepts (Kulkarni et al., 2015; Reed et al., 2015). This area has received renewed interest after the theoretical impossibility result of Locatello et al. (2019) and the identifiability proofs of Khemakhem et al. (2020) and Mita et al. (2021). A key aspect of recent weakly-supervised models is the interpretation of the grouping as a signal of similarity between datapoints (Chen & Batmanghe-

[1] `https://anonymous.4open.science/r/style-content-conditional-shift-0CC7`

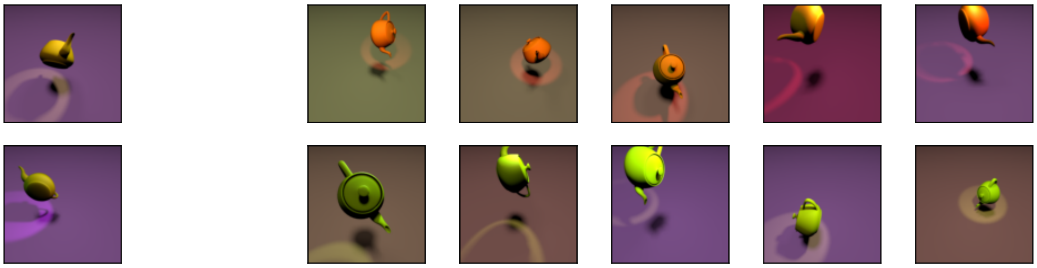


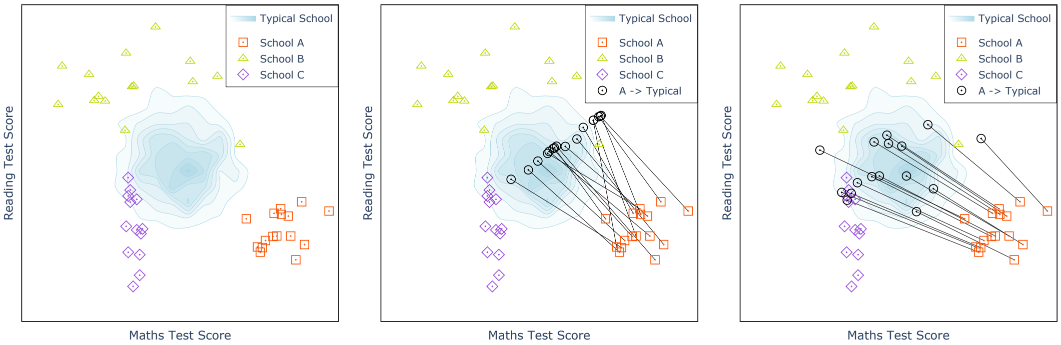
Figure 1: **Conditional shift in the `Shift3DIdent` dataset.** It appears that the objects in the two images of the leftmost column have the same colour. However, each image was generated by a different combination of object-colour, spotlight-colour: One teapot is actually orange, while the other is bright green. We can see this by looking at other examples of the same objects under different lighting conditions. The images to the right depict different views of the same corresponding object.

lich, 2020). Recently, (Kügelgen et al., 2021) have proved that style-content representations are identifiable. However, their proof relies on the assumption that the mappings between observations and latent variables are invertible, which is not the case in conditional shift.

Controlling for the Style Variable. Controlling on group-level variables is a well-established solution for dealing with confounders. However, existing methods rely on observing the confounding directly (Pearl et al., 2016). While conditioning the content encoder on the style variable is common in the areas of semi-supervised learning and fair representations (Kingma et al., 2014; Louizos et al., 2016), we are the first to apply it to unsupervised group disentanglement where explicit group labels are not available. In the field of sequence disentanglement, state-of-the-art methods (Hsu et al., 2017; Denton & Birodkar, 2017; Li & Mandt, 2018) infer the content variable (capturing shorter timescales) conditionally on the style variable (capturing longer timescales). Recent works in weakly-supervised disentanglement (Shu et al., 2020; Locatello et al., 2019; Roeder et al., 2019) also condition the content variable on the group, but their style variable is a discrete variable used for selection rather than a representation. It marks which units of the content representation are common within the group and which are free to vary. We argue that this is not sufficient to account for the variation in $p(C|X, S)$ produced by the conditional shift, so we include AdaGVAE (Locatello et al., 2020) in our evaluation for comparison (see Section 6).

Conditional Shift. This phenomenon has recently received attention in the realm of supervised learning (Liu et al., 2023). There are two main approaches to overcoming this challenge: causal methods (Scholkopf et al., 2021) and distributionally robust optimisation (Blanchet et al., 2019). Our content encoder conditioned on the style variable is a new strategy to deal with conditional shift. This problem has been studied extensively in the context of supervised learning (Zhang et al., 2013a; Gong et al., 2016). However, we are the first to explore the effect of conditional shift on unsupervised learning. Methods for mitigating the effects of conditional shift typically focus on learning domain-invariant representations (Ben-David et al., 2009). However, Zhao et al. (2019) show that learning a domain-invariant representation is not sufficient for learning a correct mapping between content variables from different groups.

Image Translation. These methods produce excellent results on high-dimensional data, but they do not use an explicit content variable that can be used for downstream tasks. Note that our variational latent posterior is different from the one used in COCO-FUNIT (Saito et al., 2020). The authors are motivated by the same limitations with existing works as we are, namely that unsupervised translation methods struggle to disentangle under conditional shift. However, because they train explicitly for translation rather than disentanglement, they arrive at a different solution than ours. When performing a translation, their approach is to condition the representation of the target group on the source image, thereby bypassing the need for an accurate content representation. This mechanism produces impressive results on image translation tasks, but it cannot be extended to models based on the GVAE which do not train explicitly for translation; in our case, there are no source and target groups in the training set. Regardless, we evaluate COCO-FUNIT on the test score dataset and show that our model outperforms it both in terms of disentanglement and translation.



(a) Test scores grouped by school. (b) Failed translation (GVAE). (c) Successful translation (ours).

Figure 2: **Our model correctly disentangles between student aptitude and school characteristics, whereas the GVAE fails.** We ask the counterfactual question "What score would student k from school A have obtained if they had attended the *typical school*?". The task is to generate a set of test scores by translating the scores from school A onto the distribution of scores from the typical school. Each line shows the translation for an individual student. Our model preserves the relative positions of the scores, whilst also capturing the distribution of the typical school.

3 Background

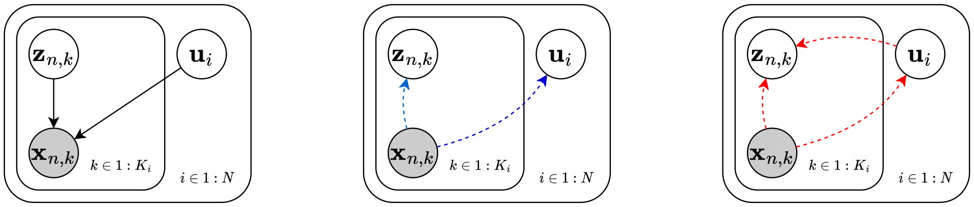
We start by formalising the problem of style-content disentanglement. Consider a training dataset of D-dimensional observations organised into N groups: $\mathbf{x}_{n,1:K_n} = \{\mathbf{x}_{n,k}\}_{k=1}^{K_n}$, $\mathbf{x}_{n,k} \in \mathbb{R}^D$, $n \in \{1:N\}$. At test time, we are given a new group of observations $\mathbf{x}_{1:K} \in \mathbb{R}^{D\times K}$. The underlying probabilistic model consists of a group of observations $\mathbf{x}_{1:K} \in \mathbb{R}^{D\times K}$, a group of content variables $\mathbf{c}_{1:K} \in \mathbb{R}^{C\times K}$, a group of target variables $\mathbf{y}_{1:K} \in \mathbb{R}^{I\times K}$, and a single style variable $\mathbf{s} \in \mathbb{R}^S$. The observations $\mathbf{x}_{1:K}$ can be seen, while the other variables are hidden. We assume that all the latent variables are independent and normally distributed $\mathbf{s} \sim \mathcal{N}(0,1)$, $\mathbf{c}_k \in \mathcal{N}(0,1)$. Each observation $\mathbf{x}_k$ depends only on the style $\mathbf{s}$ and the corresponding content $\mathbf{c}_k$ (Figure 3a). Additionally, the target variable $\mathbf{y}_k$ only depends on the corresponding content $\mathbf{c}_k$. The joint distribution factorises as $p(\mathbf{x}_{1:K}, \mathbf{y}_{1:K}, \mathbf{c}_{1:K}, \mathbf{s}) = p(\mathbf{s}) \prod_{k=1}^{K} p(\mathbf{c}_k)\, p(\mathbf{x}_k|\mathbf{c}_k, \mathbf{s})\, p(\mathbf{y}_k|\mathbf{c}_k)$.

The goal of style-content disentanglement is to learn a distribution $q_\phi(\mathbf{s}, \mathbf{c}_{1:K}|\mathbf{x}_{1:K})$ that infers the latent variables $(\mathbf{s}, \mathbf{c}_{1:K})$ conditionally on the group of observations $\mathbf{x}_{1:K}$ in a way that matches the true latent posterior $p(\mathbf{s}, \mathbf{c}_{1:K}|\mathbf{x}_{1:K})$ as closely as possible. The content variables $\mathbf{c}_{1:K}$ will then be used to train a predictor for the target variable $h(\mathbf{y}_k|\mathbf{c}_k)$. Although this final step is outside the scope of style-content disentanglement per se, we can use the accuracy of the resulting predictor h as an evaluation metric.

3.1 Group Variational Autoencoder

The conventional approach for style-content disentanglement is to learn the representation network $q_\phi(\mathbf{s}, \mathbf{c}_{1:K}|\mathbf{x}_{1:K})$ by using it as the variational latent posterior distribution in a Variational Autoencoder (Bouchacourt et al., 2018; Hosoya, 2019; Németh, 2020). In addition to the representation network, the autencoder comprises a generator $p_\theta(\mathbf{x}_k|\mathbf{s}, \mathbf{c}_k)$ decoding each observation $\mathbf{x}_k$ from the style $\mathbf{s}$ and its corresponding content $\mathbf{c}_k$. The two networks p_θ, q_ϕ are trained by maximising the Evidence Lower Bound (ELBO) (Kingma & Welling, 2014; Rezende et al., 2014) on the training data. The ELBO for a group $n \in \{1:N\}$ has the following form:

$$\mathbb{E}_{q_\phi(\mathbf{s}_n, \mathbf{c}_{n,1:K}|\mathbf{x}_{n,1:K_n})}\left[\sum_{k=1}^{K_n} \log p_\theta(\mathbf{x}_{n,k}|\mathbf{s}_n, \mathbf{c}_{n,k})\right] - \mathrm{KL}\left[q_\phi(\mathbf{s}_n, \mathbf{c}_{n,1:K}|\mathbf{x}_{n,1:K}) \,||\, p(\mathbf{s}_n) \prod_{k=1}^{K_n} p(\mathbf{c}_{n,k})\right]$$



(a) Group generative model (b) Inference model of GVAE (c) Our inference model (CxVAE)

Figure 3: Our variational distribution conditions the content variable $\mathbf{c}_k$ on the style variable $\mathbf{s}$.

When defining the variational latent posterior, virtually all existing works make the modelling assumption that the style and content variables are conditionally independent given the observations $q_\phi(\mathbf{s}, \mathbf{c}_{1:K}|\mathbf{x}_{1:K}) = q_\phi(\mathbf{s}|\mathbf{x}_{1:K}) \sum_{k=1}^{K} q_\phi(\mathbf{c}_k|\mathbf{x}_k)$ (Figure 3b).

This assumption significantly simplifies the computational architecture of the representation network by splitting it into two networks: a style encoder $q_\phi(\mathbf{s}|\mathbf{x}_{1:K})$ and a content encoder $q_\phi(\mathbf{c}_k|\mathbf{x}_k)$. However, we claim that this assumption is partly to blame for the diffiiculty of existing models to disentangle under conditional shift.

4 Context-Aware Variational Autoencoder

We propose a new inference model — the Context-Aware Variational Autoencoder (CxVAE) — that can perform style-content disentanglement on datasets with conditional shift. There are two key differences between the GVAE and our model:

Conditioning on the style. We modify the content encoder by adding the style variable $\mathbf{s}$ in the conditioning $q_\phi(\mathbf{c}_k|\mathbf{x}_k)$. This formulation reflects the factorisation of the underlying probabilistic model that does not make any independence assumptions $p(\mathbf{s}, \mathbf{c}_{1:K}|\mathbf{x}_{1:K}) = p(\mathbf{s}|\mathbf{x}_{1:K}) \sum_{k=1}^{K} p(\mathbf{c}_k|\mathbf{x}_k)$. Conditioning the content $\mathbf{c}_k$ on the style $\mathbf{s}$ represents a necessary step towards enabling the encoder q_ϕ to produce a group of content variables $\mathbf{c}_{1:K}$ that are marginally independent of one another. This is because the observation $\mathbf{x}_k$ acts as a probabilistic collider between the style $\mathbf{s}$ and the content $\mathbf{c}_k$ (Pearl et al., 2016).

Adversarial loss. In addition to the ELBO, we train the representation network q_ϕ explicitly to produce groups of content variables $\mathbf{c}_{1:K}$ that are marginally independent of one another. We train an adversarial network $t : \mathbb{R}^{C\times *} \to \{0, 1\}$ whose task is to classify whether a set of content variables $\{\mathbf{c}_1, \dots, \mathbf{c}_T\}$ are inferred from the same group. During training, the network receives as input either a set of content variables from the same group $\mathbf{c}_{n,1:K_n}$, for which it should output the value 1, or a random selection of content variables from different groups $\{\mathbf{c}_{\sigma_1}, \dots \mathbf{c}_{\sigma_T}\}$, for which it should output the value 0. The encoder network $\mathbf{q}_\phi$ is trained to reduce the accuracy of the adversarial network $\mathbf{t}$. By implication, this objective reduces the mutual information among the content variables within a group $\mathbf{c}_{1:K}$. The loss of the CxVAE is $\max_{\theta,\phi} \min_{t} \mathrm{ELBO}(\theta, \phi) + \mathrm{ADV}(\theta, \phi, t)$, where the ELBO is:

$$\mathrm{ELBO}(\theta, \phi) = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{q_\phi(\mathbf{s}_n, \mathbf{c}_{n,1:K_n}|\mathbf{x}_{n,1:K_n})} \left[\sum_{k=1}^{K_n} \log p_\theta(\mathbf{x}_{n,k}|\mathbf{s}_n, \mathbf{c}_{n,k}) \right]$$

$$-\frac{1}{N} \sum_{n=1}^{N} \mathrm{KL}\left[q_\phi(\mathbf{s}_n|\mathbf{x}_{n,1:K_n}) \,\|\, p_\theta(\mathbf{s}_n)\right] - \mathbb{E}_{q_\phi(\mathbf{s}_n|\mathbf{x}_{n,1:K_n})} \left[\sum_{k=1}^{K_n} \mathrm{KL}\left[q_\phi(\mathbf{c}_{n,k}|\mathbf{x}_{n,k}, \mathbf{s}_n) \,\|\, p_\theta(\mathbf{c}_{n,k})\right] \right]$$

5 Evaluation

We show that our CxVAE produces a considerable improvement in disentanglement over competing style-content disentanglement methods. We also show that this improvement is proportional to the strength of the conditional shift effect in the dataset.

Model Setup. We compare our conditional CxVAE with the state-of-the-art in group disentanglement, namely the GVAE (Hosoya, 2019; Bouchacourt et al., 2018), COCO-FUNIT (Saito et al., 2020), and AdaGVAE (Locatello et al., 2019). As in Hosoya (2019), the group encoder is applied to each datapoint in the group and then all the outputs are averaged.

For all experiments, our CxVAE will be a modified GVAE such that the style variable $\mathbf{s}$ is concatenated with the observation $\mathbf{x}_k$ and fed into the content encoder in order to compute the content variable $\mathbf{c}_k$. For sampling the variational latent posteriors, we use the standard reparametrisation trick. We use an Adam optimiser with learning rate of 1e-4 with $\beta_1 = 0.9, \beta_2 = 0.5$. For the `Shift3DIdent` dataset, we implement all networks (encoders and decoders) as convolutional nets with 4 hidden layers and 64 filters each. Both latent variables have 16 latent dimensions. For the `TestScores` dataset, we use MLPs with 3 hidden layer of 32 activations each. The style variable will have 4 dimensions and the content variable will have 2 dimension.

We train each model for 64 epochs, and use the last 10 epochs for evaluation. Additionally, we run the experiment for 100 different random seeds initialisations, both for the data generating process and the networks. Confidence intervals are computed by resampling train-test splits, weight initialisations and sampling seeds. We use the same 100 seeds in each model. This gives 1000 measurements to plot in Table 1.

5.1 Datasets

`TestScores`. Consider the task of fair comparisons between students attending different schools based on their standardised test scores in maths and reading (Braun et al., 2006). The typical assumption in the literature is that each school has a similar distribution of aptitude among its students, so our goal is to learn disentangled student-level (content) and school-level (style) representations of the scores. The content representation $\mathbf{c}_{n,k}$ should reflect the aptitude of student k from school n independently of which school the student had attended. Additionally, we treat the content $\mathbf{c}_{n,k}$ as the ground-truth target variable $\mathbf{y}_{n,k}$ and use it for evaluating the quality of the content representation. This analysis is crucial for university admission boards aiming to judge students based on their aptitude regardless of the socio-economic circumstances associated with attending one school or another, such as affluence, location, or curriculum (Raudenbush & Willms, 1995; Braun et al., 2006).

We generate our `TestScores` using the classic "varying intercept, varying slope" mixed-effects model (Laird & Ware, 1982; Pinheiro & Bates, 2001; Gelman & Hill, 2006). This is a well-established approach for modelling student scores $\mathbf{x}_{n,k}$ as a function of individual aptitude $\alpha_{n,k}$ and school-level characteristics (β_n, γ_n), e.g., affluence, curriculum, or location (Raudenbush & Willms, 1995; Braun et al., 2006). We choose this model for its simplicity and for the wide variety of phenomena to which it can be applied. All the scores and factors are 2-dimensional vectors, with one component for the maths score and another for the reading score. $\mathbf{x}_{n,k} = \beta_n - \gamma_n \odot \alpha_{n,k} + \epsilon_{n,k}$ is the score of student k from school n. $\alpha_{n,k} \sim \mathcal{N}(0, I_2)$ is the aptitude of student k in school n. $\beta_n \sim \mathcal{N}(0, I_2)$ is the mean score in school n. $\gamma_n \sim \text{Exp}(1)$ is the standard deviation of scores in school n. $\epsilon_{n,k} \sim \mathcal{N}(0, 0.1 * I_2)$ is a per-student error term. For the evaluation procedure, we use the above model to generate $N = 32{,}768$ values for (β_n, γ_n). For each school n, we generate $M = 128$ values for $\alpha_{n,k}$. We then randomly select half of the schools to assemble a training dataset with 2,097,152 scores split across 16,384 schools. We take the other half of schools to create the holdout dataset, so that every testing school and student are unseen during training.

Looking at the data (Figure 2a), it is clear that we are dealing with the conditional shift scenario. The same reading score could be obtained by either a high-achieving student from school C or a low-achieving student from school B. Inferring the aptitude of the student requires knowing the distribution of scores within each school. In other words, the distribution of aptitude $\mathbf{c}$ given the score $p(\mathbf{c}_{n,k}|\mathbf{x}_{n,k})$ changes from one school n to another.

`Shift3DIdent`. This dataset comprises images from the 3DIdent dataset (Zimmermann et al., 2021) depicting teapots of different colours being lit by spotlights of different colours. Within a group, the images have the same spotlight colour and only the colour of the teapot varies. The goal is for the style representation to encode the spotlight colour and for the content representation to encode the object colour. This goal is useful for many real-world applications, such as object recognition. Once we have separated the group-level variation from the instance-level variation, we can use the content representation as a low-level feature on which to train a classifier that predicts the colour of

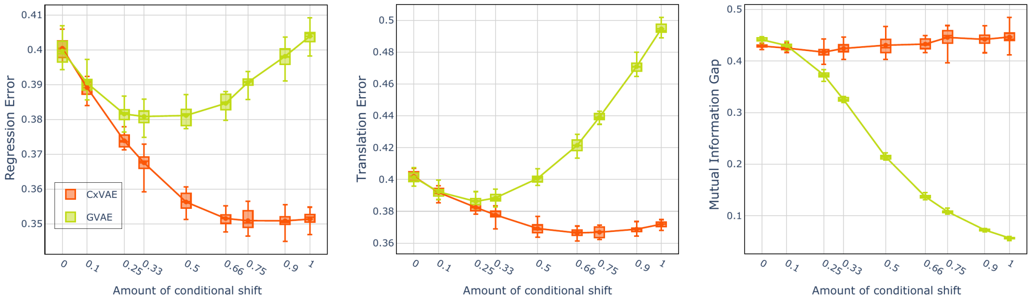


Figure 4: **Conditional shift in the dataset causes the relative performance gain of our CxVAE over the GVAE.** We show performance on datasets generated with different values of the λ hyper-parameter, controlling the amount of conditional shift. For low values of λ the conditional distribution of student aptitude given a score changed very little from one group to another, and so CxVAE and GVAE perform equally well. However, as λ increases, the GVAE fails to disentangle.

the object. With better disentanglement the predictor would have higher accuracy at identifying the true colour of the object.

This is a challenging problem because the data exhibits conditional shift. The same exact image could have been generated by different combination of spotlight colour and object colour, as can be seen in Figure 1. This makes it difficult to identify the true colour of the object by just looking at one single image. Indeed, all existing methods infer the content variable using only one image and fail to learn disentangled representations on this dataset.

To generate the dataset, we select $N = 4{,}096$ spotlight colours (styles) and, for each style, generate $K = 16$ object colours (contents). Half of the groups are kept for training, and half for testing. We combine each content with the corresponding style and generate the observation $\mathbf{x}$. We treat the value of the object colour as the target variable $\mathbf{y}$.

5.2 Evaluation Metrics

Target Variable Regression. After inferring the content variables $\mathbf{c}$ for the entire testing set, we use half of them as training inputs for a model to predict the target variable $h(\mathbf{y}|\mathbf{c})$. We use the other half to evaluate this predictor by taking the average mean-squared error between the prediction and the ground-truth $\mathbf{y}$. This error will be our first measure of disentanglement. It reflects how much information about the target y is contained in the content $\mathbf{c}$.

Mutual Information Gap. We use the Mutual Information Gap (Chen et al., 2018) to measure the quality of the disentanglement. We measure empirically the amount of mutual information between the inferred latent variables $\mathbf{s}, \mathbf{c}$ and the ground-truth style factor $\mathbf{s}'$. Consequently, the goal is to have maximal mutual information between the style variable $\mathbf{s}$ and the ground-truth style $\mathbf{s}'$, and minimal mutual information between the content variables $\mathbf{c}$ and the ground-truth $\mathbf{s}'$. The gap between the two (normalised with the entropy of the ground-truth factors) is the metric of disentanglement $\mathrm{MIG} = \frac{1}{H(\mathbf{s}')}(I(\mathbf{s};\mathbf{s}') - I(\mathbf{c};\mathbf{s}'))$. Since the data-generating process is known, the mutual information between the inferred style variable and the ground-truth style variable $I(\mathbf{s};\mathbf{s}')$ is straightforward to implement by following the approach from Chen et al. (2018). We measure only the mutual information between the ground-truth group factor $\mathbf{s}'$ and the latent variables because, as pointed out by Németh (2020), the common failure case we are trying to guard against in style-content disentanglement is that the content variables $\mathbf{c}$ might learn information belonging to the ground-truth style factor $\mathbf{s}'$.

Translation Between Styles. We measure how well the learned representations can answer the question “What would the score of student k from school n have been if they had attended the typical school (the school with scores distributed according to $\mathcal{N}(0, I_2)$)?”. This problem, also known as translation, is a commonly used downstream task for disentangled representations (Tenenbaum & Freeman, 2000). We translate the score of student k to the typical school and then take the mean

Table 1: Our CxVAE model outperforms existing methods at disentanglement on the `Shift3DIdent` and `TestScores` datasets. The competing models are: GVAE (Hosoya, 2019), AdaGVAE (Locatello et al., 2020), and COCO-FUNIT (Saito et al., 2020). The improvement in scores brought by our model is greater than the 95% confidence interval around each score. Confidence intervals are computed by resampling train-test splits, weight initialisations and sampling seeds.

	Shift3DIdent			**TestScores**		
MODEL	RGRSN ↓	MIG ↑	TRNSL ↓	RGRSN ↓	MIG ↑	TRNSL ↓
GVAE	0.08 ± 0.02	0.09 ± 0.06	0.23 ± 0.18	0.41 ± 0.01	0.04 ± 0.02	0.49 ± 0.01
AdaGVAE	0.09 ± 0.05	0.39 ± 0.11	0.15 ± 0.05	0.41 ± 0.01	0.05 ± 0.02	0.49 ± 0.01
CFUNIT	**0.05** ± 0.01	0.13 ± 0.04	0.13 ± 0.06	0.40 ± 0.02	0.04 ± 0.01	0.49 ± 0.02
CxVAE	**0.05** ± 0.03	**0.72** ± 0.06	**0.08** ± 0.05	**0.35** ± 0.02	**0.44** ± 0.08	**0.37** ± 0.02

squared error against the ground-truth translation, which is generated when the data is generated using the ground-truth generative factors. For our dataset, the correct translation corresponds to the Earth-Mover distance between the multivariate normal distributions of scores in each school (Knott & Smith, 1984).

In order to obtain the translation, we first infer the content variable of student k from school n and the style variable of the typical school. We then feed the two variables to the decoder. For evaluation, we translate all the scores from each school n and then measure the distance between the predicted translation and the ground-truth translation. We compute the total error as an average over all the translation errors.

Translation is a well-established downstream task for evaluating disentanglement (Tenenbaum & Freeman, 2000). In the case of the `TestScores` dataset, translation corresponds to the counterfactual question "What score would student k from school A have obtained if they had attended the *typical school* (i.e., a school whose scores are distributed according to $\mathcal{N}(0, I_2)$)?" We can also use translation as an additional qualitative comparison between CxVAE and other group disentanglement methods, which can be seen in Figure 2.

6 Results

We show that our CxVAE is able to style-content disentangle in a setting where conditional shift produces ambiguous observations. The results in Table 1 show that our model, CxVAE, produces representations that disentangle between spotlight colour and object colour much more than a representative selection of existing models: GVAE, AdaGVAE, and COCO-FUNIT. Our model's MIG score is much higher than the one produced by competing models and the performance gap is greater than the 95% confidence interval for any one of these models.

Our CxVAE produces considerable improvements over the competing methods in terms of fitting the holdout set, disentangled representations, translation accuracy, and predicting the generative factors (Table 1). While the scores of the existing methods cluster together, the gap between them and CxVAE is larger than the 95% confidence interval of any method.

7 Conditional Shift Causes the Gap in Performance

By modifying the data-generating process (Equation **??**), we show that conditional shift explains the increased performance of CxVAE. We insert a hyper-parameter λ to control the strength of the conditional shift; $\lambda = 1$ means the conditional shift stays the same as in the previous experiment, while $\lambda = 0$ means there is no conditional shift.

Consider the case where the maths score only depends on the school and the reading score only depends on the student. In this situation, the two generative factors can be easily disentangled since you can infer the student aptitude from the reading score and the school profile from the maths score. We use this as an extreme case of lack of conditional shift and insert a hyper-parameter λ in our data-generating process that will continuously move between this case and the original case. Our modified data-generating process is $\mathbf{x}_{n,k} = \begin{bmatrix}\lambda\\1\end{bmatrix} \odot \beta_n + \begin{bmatrix}1\\\lambda\end{bmatrix} \odot \alpha_{n,k} \odot \left(\gamma_n \hat{\cdot} \begin{bmatrix}\lambda\\1\end{bmatrix}\right) + \epsilon_{n,k}$, where $\hat{\cdot}$ denotes the elementwise power operation and the generative factors $(\alpha_{n,k}, \beta_n, \gamma_n, \epsilon_{n,k})$ are sampled as before. This model has no conditional shift when $\lambda = 0$ because each ground-truth factor controls a separate component of the data. Inferring the student aptitude requires only the reading score and can ignore the school characteristics. When $\lambda = 1$, the problem exhibits conditional shift in exactly the same way as before.

We hypothesise that the conditional shift causes the performance gap between CxVAE and other group disentanglement methods. If our hypothesis is correct, then the gap should decrease as λ approaches 0. The measurements displayed in Figure 4 confirm our expectations. For low values of λ the performance of our CxVAE is evenly matched to the GVAE. As λ increases, CxVAE metrics remain stable while GVAE performance decreases substantially. It is clear that the degree of confounding in the dataset explains the performance gain that we see in CxVAE.

8 Conclusions

In this work, we show empirically that conditioning the content encoder on the style variable produces style-content disentangled representations on datasets with conditional shift. We also show that the strength of the conditional shift effect in the data-generating process determines the amount of improvement that our model brings over other group disentanglement methods. Our evaluation is run on two important downstream tasks for style-content disentanglement: One is the problem of inferring student aptitudes from test scores grouped by school, and another is identifying the colours of objects viewed under different lighting conditions.

9 Reproducibility Statement

Details for reproducing the experiments are presented in Section 5 and on the anonymised GitHub repo: `https://anonymous.4open.science/r/style-content-conditional-shift-0CC7`.

References

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F. C., and Vaughan, J. W. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2009.

Blanchet, J., Kang, Y., and Murthy, K. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, September 2019. ISSN 0021-9002, 1475-6072. doi: 10.1017/jpr.2019.49.

Bouchacourt, D., Tomioka, R., and Nowozin, S. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *AAAI Conference on Artificial Intelligence*, 2018.

Braun, H., Jenkins, F., and Grigg, W. S. Comparing private schools and public schools using hierarchical linear modeling. *U.S. Department of Education, National Center for Education Statistics, Institute of Educational Sciences*, 2006.

Chen, J. and Batmanghelich, K. Weakly supervised disentanglement by pairwise similarities. *AAAI Conference on Artificial Intelligence*, 2020.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018.

Denton, E. L. and Birodkar, V. Unsupervised learning of disentangled representations from video. *ArXiv*, abs/1705.10915, 2017.

Gelman, A. and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

Gondal, M. W., Wüthrich, M., Miladinovic, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J. B., Bachem, O., Schölkopf, B., and Bauer, S. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In *NeurIPS*, 2019.

Gong, M., Zhang, K., Liu, T., Tao, D., Glymour, C., and Schölkopf, B. Domain adaptation with conditional transferable components. *JMLR workshop and conference proceedings*, 48:2839–2848, 2016.

Higgins, I., Matthey, L., Pal, A., Burgess, C. P., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Hosoya, H. Group-based learning of disentangled representations with generalizability for novel contents. In *IJCAI*, 2019.

Hsu, W.-N., Zhang, Y., and Glass, J. R. Unsupervised learning of disentangled and interpretable representations from sequential data. In *NIPS*, 2017.

Khemakhem, I., Kingma, D. P., and Hyvärinen, A. Variational autoencoders and nonlinear ica: A unifying framework. *ArXiv*, abs/1907.04809, 2020.

Kim, H. and Mnih, A. Disentangling by factorising. *ArXiv*, abs/1802.05983, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2014.

Kingma, D. P., Mohamed, S., Rezende, D. J., and Welling, M. Semi-supervised learning with deep generative models. *NeurIPS*, 2014.

Knott, M. and Smith, C. S. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49, 1984.

Kügelgen, J. V., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M., and Locatello, F. Self-supervised learning with data augmentations provably isolates content from style. 2021.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. B. Deep convolutional inverse graphics network. In *NIPS*, 2015.

Laird, N. M. and Ware, J. H. Random-effects models for longitudinal data. *Biometrics*, 38(4): 963–974, 1982.

LeCun, Y., Huang, F. J., and Bottou, L. Learning methods for generic object recognition with invariance to pose and lighting. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2:II–104 Vol.2, 2004.

Li, Y. and Mandt, S. Disentangled sequential autoencoder. In *ICML*, 2018.

Liu, J., Wang, T., Cui, P., and Namkoong, H. On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, November 2023.

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Locatello, F., Poole, B., Rätsch, G., Scholkopf, B., Bachem, O., and Tschannen, M. Weakly-supervised disentanglement without compromises. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. S. The variational fair autoencoder. *CoRR*, abs/1511.00830, 2016.

Mita, G., Filippone, M., and Michiardi, P. An identifiable double vae for disentangled representations. In *ICML*, 2021.

Németh, J. Adversarial disentanglement with grouped observations. In *AAAI*, 2020.

Pavlou, M., Ambler, G., Seaman, S. R., Guttmann, O., Elliott, P., King, M., and Omar, R. Z. How to develop a more accurate risk prediction model when there are few events. *BMJ*, 351:h3868, August 2015. ISSN 1756-1833. doi: 10.1136/bmj.h3868.

Pearl, J., Glymour, M., and Jewell, N. P. *Causal Inference in Statistics: A Primer*. Wiley, Chichester, West Sussex, 2016. ISBN 978-1-119-18684-7.

Pinheiro, J. C. and Bates, D. M. Mixed-effects models in s and s-plus. *Technometrics*, 43:113 – 114, 2001.

Raudenbush, S. and Willms, J. D. The estimation of school effects. *Journal of Educational Statistics*, 20:307 – 335, 1995.

Reed, S. E., Zhang, Y., Zhang, Y., and Lee, H. Deep visual analogy-making. In *NIPS*, 2015.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

Roeder, G., Grant, P. K., Phillips, A., Dalchau, N., and Meeds, E. Efficient amortised bayesian inference for hierarchical and nonlinear dynamical systems. *ArXiv*, abs/1905.12090, 2019.

Saito, K., Saenko, K., and Liu, M.-Y. Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. *ECCV*, 2020.

Scholkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109(5):612–634, May 2021. ISSN 0018-9219, 1558-2256. doi: 10.1109/JPROC.2021.3058954.

Shu, R., Chen, Y., Kumar, A., Ermon, S., and Poole, B. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020.

Tenenbaum, J. B. and Freeman, W. T. Separating style and content with bilinear models. *Neural Computation*, 12:1247–1283, 2000.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain adaptation under target and conditional shift. In *ICML*, 2013a.

Zhang, K., Schölkopf, B., Muandet, K., and Wang, Z. Domain Adaptation under Target and Conditional Shift. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 819–827. PMLR, May 2013b.

Zhao, H., des Combes, R. T., Zhang, K., and Gordon, G. J. On learning invariant representations for domain adaptation. In *ICML*, 2019.

Zimmermann, R. S., Sharma, Y., Schneider, S., Bethge, M., and Brendel, W. Contrastive learning inverts the data generating process. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 12979–12990, 2021.