Do VAEs Preserve Causal Relationships Present in Data in the Latent Space?

Petra Poklukar*[1], Michael C. Welle*[1]

Abstract— **Being able to preserve known causal structures in generative frameworks like Variational Auto-Encoders (VAEs) would enable the generation of constrained datasets as well as enhance the understanding and explainability of such models. In this project we compare the structure of the latent space of models trained on a causal and non-causal version of limited d-sprite dataset. We define a novel measure called *Interventional score* inspired by the causality framework which we benchmark with two established disentanglement scores. We show that the current tools are insufficient to include causal graphs into the VAE framework and demonstrate the difficulty of their evaluation. We conclude the project with potential future direction and pose open questions to the community.**

I. INTRODUCTION

Identifying underlying generative factors of a dataset is of great interest for the Robotics and Machine Learning (ML) communities. As big datasets and computational power became more available the resurgence of Neural Networks (NNs) contributed to big strides in these fields in the recent years. While a single layer feed-forward NN can in theory approximate any function [1], stacking them into a Deep Neural Network (DNN) enabled the communities to come much closer to this goal in a practical sense as before. The core difference between Deep Learning approaches and more traditional ones is the absence of handcrafted features. A DNN is guided only by a loss function[1] to learn the features from the data instead of relying on hand engineered features that are task specific and might miss important information present in the data. The downside of this data-driven approach is the risk of learning spurious features, the problem known as the ”Clever Hans” effect, which can lead to overconfidence and consequently to a decrease in the performance of DNNs [2]. To this end, an area of interest in the recent years in the ML community has been improving the explainability and interpretability of DNNs [3], [4], [5], [6]. One step towards this goal is to measure the *disentanglement* of the features learned by DNNS. The core idea is to associate each underlying generative factor of a given dataset with a set of components of the learned features. This however is only fully possible if the underlying factors of generation are independent of each other, a requirement that is very seldom fulfilled in real data.

A research field that specifically investigates the underlying data dependencies and ways to interfere, validate and exploit them is *causality*. In particular, it studies the *cause-effect* relationships between pairs of random variables which are described with Structural Causal Models (SCMs) [7]. An SCM induces a *causal graph* where parent nodes are the causes of the descendent nodes. In order to improve the disentanglement analysis one should therefore integrate the causal information encoded in the given dataset into DNNs. Note that the problem of inferring the causal structure from the data lies outside the scope of this project.

Surprisingly, current literature has very few works that consider causality as a potential new tool for analysing disentanglement tasks. With this project we take a first step towards this direction and investigate if a Variational Autoencoder (VAE) [8], [9] is able to preserve the causal structure present in the data in its latent space. In detail our contributions are:

- Two novel image datasets, called *Causal* and *Non-Causal*, generated using two SCMs that differ in the number of cause variables. As a basis, we use a subset of the d-Sprite 2D dataset [10] which is commonly used to evaluate disentanglement.
- A novel investigation of whether or not VAEs captures the underlying causal structure of the *Causal* and *NonCausal* datasets in the latent space. We define a new measure, which we call *Interventional score*, which uses interventions in the latent space combined with the Maximum Mean Discrepancy (MMD) [11] to determine if the true causal graph of the input dataset has been preserved in the encoding process.
- We evaluate our approach on VAEs trained with five different latent dimensions on both Causal and Non-Causal datasets and compare it with the two different disentanglement scores introduced by [12] and [13].
- We identify and discuss the possible future directions for this line of work.

II. RELATED WORK

The idea of disentanglement has quite a long history and dates at least back to work on basic autoencoders [14] but gained a lot of interest with the rise of representation learning [15] with DNNs. In this project however we will mostly focus on the work of Kim and Mnih, ”Disentangling by Factorising”[12] and ”A Framework for the Quantitative Evaluation of Disentangled Representations” [13] by Eastwood and Williams.

The core idea in [12] is to take the ELBO objective from the β-VAE:

$$\mathscr{L}_{vae}(x) = E_{z\sim q(z|x)}[\log p(x|z)] + \beta \cdot D_{KL}(q(z|x)||p(z))$$

and further decompose the KL term [16] into:

$$E_{p_{data}(x)}[D_{KL}(q(z|x)||p(z))] = I(x;z) + D_{KL}(q(z)||q(z))$$

[1] Petra Poklukar and Michael C. Welle are part of the Causality Girls research group

* This authors are equally attractive ;)

[1] and a few hundred hyper-parameters

where $I(x;z)$ is the mutual information between an input x and its latent encoding z. The reasoning behind this decomposition is that penalising $I(x;z)$ to the same amount (β) as $D_{KL}(q(z|x)||p(z))$ is counter productive for achieving better disentanglement.

The authors therefore propose a disentanglement measure where a change in one latent dimension z_i corresponds to exactly one factor of variation, which is what we use in this project as well (section V). The metric is the error rate of a majority vote classifier trained on data generated with one factor held constant.

In contrast, the authors of the framework in [13] opt to use a lasso and random forest regressor to calculate three values that give an idea about the learned representation: Disentanglement, Completeness and Informativeness. Disentanglement gives the degree to which a representation disentangles the underlying generative factors. Completeness measures how much of the underlying factors is captured in the representation, and Informativeness represents the amount of information a representation captures of the generative factors. It is important to note that both frameworks assume that the underlying generative factors are *known* and are *independent* of each other.

Causality is a well established field that has long existed outside the interest sphere of representation learning ([17],[18],[19]). Note that in ML community the word *causal* is often used to mean *temporal dependency* [20] instead of the SCM frameworks as defined in [21], [7] and used in this work.

Trying to bridge this gap, the work presented in [22] postulates that some of the causal factors could be discovered if an agent (a learner) is interacting with its environment. By training a policy in tandem with learning the features, an additional loss term is added to the VAE training framework which is based on the assumption that different policies correspond to disentangled features as they can be independently manipulated\changed by the agent.

In [23] causality is integrated into the evaluation of a deep representation learning framework. The authors propose to see the generative factors as a causal graph and define a disentangled causal process, where conditioning on confounding factors makes the generative factors independent of each other. Since the generative factors are also a valid adjustment set as defined in [21], interventions are performed on groups of these factors. This results in a Interventional Robustness Score that captures how much the change intervention on one group of factors influences the latent encoding of the other factors.

III. Causal and Non-Causal d-Sprite Datasets

We built our Causal and Non-Causal datasets for causal evaluation of VAE's latent space on the basis of the d-Sprite[10] dataset which is commonly used for analysing disentanglement. It consists of 737280 black and white images generated with 5 independent generative factors:

- Shape: square, ellipse, heart
- Scale: 6 values linearly spaced in [0.5, 1]
- Orientation: 40 values in [0, 2π]
- Position X: 32 values in [0, 1]
- Position Y: 32 values in [0, 1]

Our datasets are subsets of d-Sprite consisting only of the shape type hearts as it is not rotationally invariant. We fixed the scale factor to 1 and kept the orientation O, X position and Y position as underlying generative factors.

The Causal and NonCausal datasets are generated with SCMs (1) and (2), respectively, and consist of 100.000 images which we split into 85% training and 15% test sets. The difference between these datasets is in the contribution of the orientation factor O to the generation of an image I. In the Causal dataset, O is dependent on the position factors X and Y, while the three factors (X,Y,O) are independent of each other in the NonCausal dataset. The respective causal graphs are shown in Figure 1.

We denote by U the exogenous variables, or independent generative factors, and by V the endogenous variables that are descendants of U. The functions f represent the edges in the causal graph and assign a value to each endogenous variable V given the values of the rest of the variables in the model. The random variables X and Y have categorical distributions with 32 categories, while O in the NonCausal dataset is categorically distributed with 39 categories.

$$\text{Causal d-sprite } SCM: \tag{1}$$

$$U = \{X,Y\}, \qquad V = \{O,I\}, \qquad F = \{f_o, f_I\}$$

$$f_O : O = \left\lfloor \frac{X \cdot Y}{961} \cdot 39 \right\rfloor$$

$$f_I : I = f(X,Y,O)$$

We generate images of the Causal dataset by uniformly sampling X and Y and following the Causal-SCM (1). Images of the NonCausal dataset are produced in a similar fashion, by uniformly sampling X, Y, O and following the NonCausal-SCM (2). The size of the images is in both case fixed to 64×64.

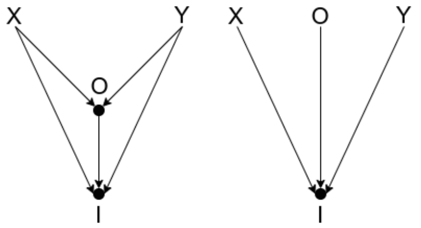

Fig. 1. Left: Causal graph from SCM 1 for Causal dSprite dataset. Right: Causal graph from SCM 2 for Non-Causal dSprite dataset.

$$\text{Non-Causal d-sprite } SCM: \tag{2}$$

$$U = \{X,Y,O\}, \qquad V = \{I\}, \qquad F = \{f_I\}$$

$$f_I : I = f(X,Y,O)$$

Examples of the generated images from both datasets are shown in Figure 2. Note that as opposed to the Causal images (left), the rotations of the hearts in the NonCausal images (right) are not linked to their positions.

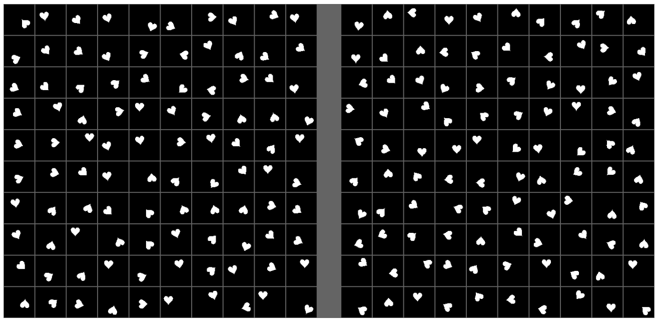

Fig. 2. Example images produced with the Causal data SCM 1 (left) and non causal data SCM 2.

IV. VAE

We deployed 5 VAEs having $2,3,4,6$ and 10 dimensional latent spaces. For the encoder and decoder networks we used $2D$ convolutional layers, ReLU activation function and dropout layers with probability $p=0.2$. The exact architecture details can be found in the code available on Causal Girl's GitHub[2]. For the specifics of the VAEs we refer the reader to [8], [9]. Note that we use the vanilla VAE framework and not any of the versions used for disentanglement, such as FactorVAE [12] or β-VAE [24].

We denote by *C-ld-n* and *NC-ld-n* the VAEs trained on the Causal and Noncausal datasets, respectively, with n being the model's latent space dimension. The *C-ld-n* models were trained for 1000 epochs, while *NC-ld-n* were trained for 4500 epochs because of the difference in complexity of the Causal and NonCausal datasets. The learning rate was fixed to $1e-4$.

Examples of reconstructions of Causal and NonCausal dataset are shown in Figures 3 and 4, respectively. It can be clearly seen that the quality of the reconstructions produced by noncausal models *NC-ld-n* is significantly worse than that of the causal models *C-ld-n*. This is likely due to NonCausal dataset being more complex. As the orientation is not related to the position of the hearts, the model learns to reconstruct either a circle or a triangle. The former appears when the latent dimension is set to 2 which could indicate that the model does not have enough capacity to disentangle the 3 independent generative factors.

V. Experiments

We investigated if the different versions of the VAE models are able to capture the causal dependency present in the data in their latent spaces. We compared the obtained results across different latent dimensions as well as across Causal and NonCausal datasets. Using disentanglement scores from [12] and [13] we first identified which latent dimension corresponds to which generative factor. We then used intervention on the latent dimensions to determine if the structure of the SCM was preserved in the latent space.

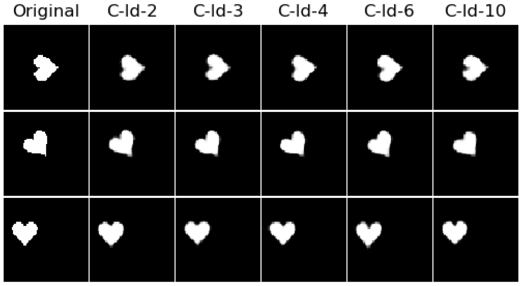

Fig. 3. Example of reconstructions of Causal dataset images for VAEs with latent dimensions (left-to-right) 2,3,4,6 and 10.

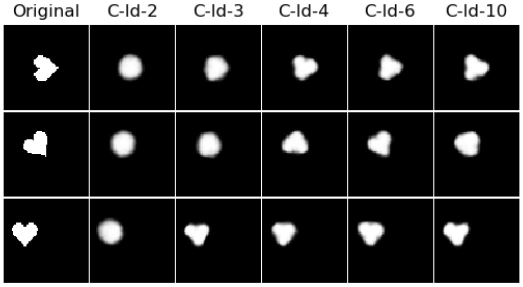

Fig. 4. Example of reconstructions of NonCausal dataset images for VAEs with latent dimensions (left-to-right) 2,3,4,6 and 10.

A. Disentagelment scores by Kim and Mnih 2019

We employed the disentanglement score proposed in [12]: ”Choose a factor k; generate data with this factor fixed but all other factors varying randomly; obtain their representations; normalise each dimension by its empirical standard deviation over the full data (or a large enough random subset); take the empirical variance in each dimension of these normalised representations. Then the index of the dimension with the lowest variance and the target index k provide one training input/output example for the classifier ...”. The final disentanglement score then corresponds to the accuracy of the classifier.

Figure 5 shows this disentanglement score [12] (Y-axis) computed using the causal models *C-ld-n* against the intermediate snapshots of the models saved during training (X-axis). Similarly, Figure 6 shows the disentanglement score computed using the noncausal models *NC-ld-n*.

B. Disentanglement, Completeness, and Informativeness by Eastwood and Williams 2018

In this section we show the Hinton diagrams of the disentanglement score as well as the informativeness and completeness scores from [13]. We provide a short summary of these measures in the following but refer the reader to [13] for further details. We denote by z_i the i-th component of a latent representation z.

[2] https://github.com/petrapoklukar/causal_disentanglement_miniproject

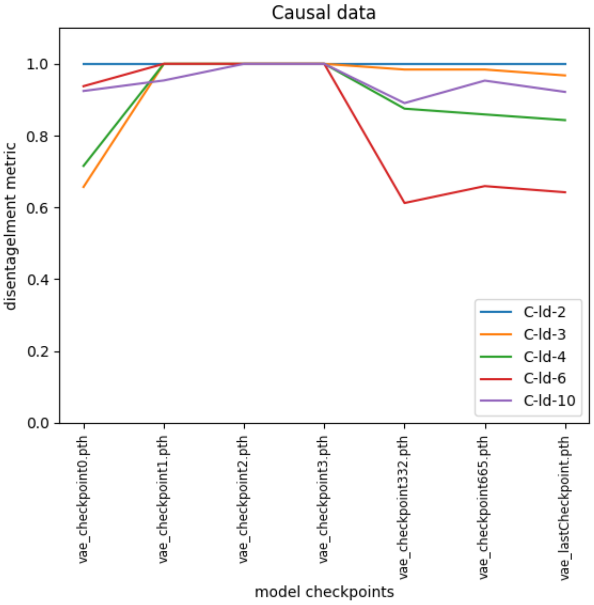

Fig. 5. Disentanglement score from [12] over several checkpoints saved during training for all 5 VAEs on the Causal dataset.

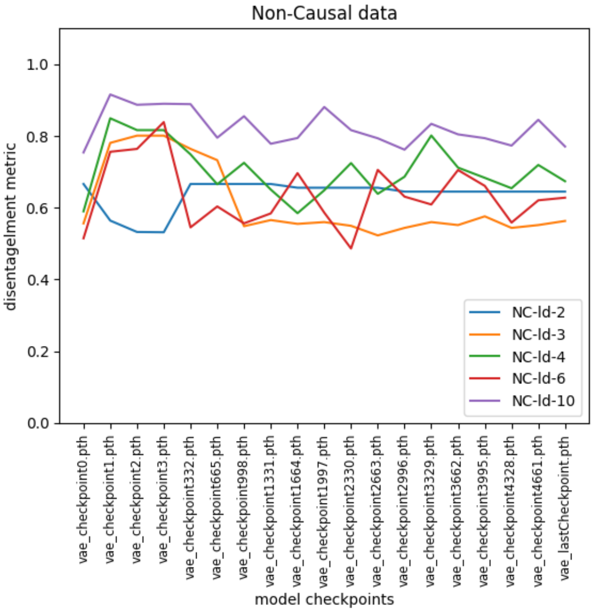

Fig. 6. Disentanglement score from [12] over several checkpoints save during training for all 5 VAEs on the Non-Causal dataset.

Disentanglement is a value between 0 and 1 indicating to which extend each latent dimension z_i captures at most one generative factor g_j. This means that if z_i is instrumental to predict a single generative factor g_j the score will be equal to 1. If z_i contributes equally to the predictions of all generative factors or is irrelevant to the prediction altogether, the score will be 0. The disentanglement therefore measures how well the latent representation unties the underling factors of generation. It can be visualised using Hinton diagrams where each row represents a latent dimension and each column corresponds to a generative factor. A larger disentanglement score for a given pair of latent dimension and generative factor is visualised with a larger white square and represents a better result (↑). Note that the a single value score per latent dimension can be obtained as a weighted average of the scores in the Hinton diagrams, the exact procedure for this is described in [13].

Completeness is a value between 0 and 1 indicating how well a latent dimension z_i captures an underlying factor g_j. If a single z_i contributes to predicting a single g_j, the score is 1 (complete). If every z_i equally contributes to the prediction of a single g_j, the score is 0 (maximally overcomplete). This is visualised in the columns of the Hinton diagrams where higher numbers are better (↑).

Informativeness is a measure that captures the amount of information that a representation contains about the underlying factors of generation. It is quantified by a prediction error using a lasso regressor. Therefore, the lower number (lower error) the better (↓).

1) Causal Dataset: Figure 7 shows the Hinton diagram for the *C-ld-n* models trained on the Causal dataset. The corresponding Disentanglement, Completeness and Informativeness scores are shown in Tables I and II.

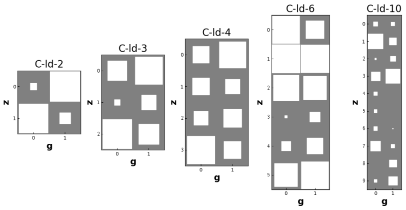

Fig. 7. Hinton diagram for the VAEs trained on the Causal dataset.

model\ld	z_0	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	Avg.
C-ld-2	**0.72**	**0.44**	-								**0.58**
C-ld-3	0.09	0.42	0.09	-							0.12
C-ld-4	0.14	0.02	0.05	0.16	-						0.11
C-ld-6	0.12	0.00	0.01	0.69	0.20	0.02	-				0.05
C-ld-10	0.02	0.30	0.54	0.14	1.00	1.00	0.83	0.50	1.00	0.54	0.36

TABLE I

THE AGGREGATED DISENTANGLEMENT SCORE (↑) PER LATENT DIMENSION COMPUTED AS IN [13] ON THE CAUSAL DATASET.

2) Non-Causal Dataset: Figure 8 shows the Hinton diagram for the VAEs trained on the NonCausal dataset. The corresponding Disentanglement, Completeness and Informativeness scores are shown in Tables III and IV.

C. Interventional score using Maximum Mean Discrepancy

In this section, we present the main contribution of our paper. We define a novel *Interventional Score* using which we can 1) test if the ground truth SCM was preserved in the latent space, and 2) calculate the extent to which each

Completeness ↑					Informativness ↓		
model\gen	g_0	g_1	Avg.		g_0	g_1	Avg.
C-ld-2	**0.71**	**0.46**	**0.59**		0.48	0.42	0.45
C-ld-3	0.34	0.11	0.22		0.29	0.41	0.35
C-ld-4	0.11	0.08	0.10		0.36	0.46	0.41
C-ld-6	0.17	0.09	0.13		0.35	0.50	0.43
C-ld-10	0.35	0.32	0.34		**0.27**	**0.35**	**0.31**

TABLE II

COMPLETENESS (↑) AND INFORMATIVENESS (↓) SCORES [13] ON THE CAUSAL DATASET.

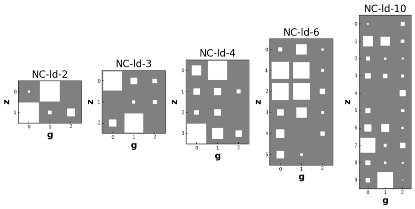

Fig. 8. Hinton diagram for the Non-Causal dataset

model\ld	z_0	z_1	z_2	z_3	z_4	z_5	z_6	z_7	z_8	z_9	Avg.
NC-ld-2	**0.96**	**0.57**	-								**0.75**
NC-ld-3	0.55	0.40	0.67	-							0.60
NC-ld-4	0.54	0.10	0.43	0.32	-						0.39
NC-ld-6	0.58	0.32	0.24	0.34	0.54	0.75	-				0.34
NC-ld-10	0.72	0.22	0.23	0.13	1.00	0.48	0.26	0.59	0.34	0.77	0.51

TABLE III

THE AGGREGATED DISENTANGLEMENT SCORE (↑) PER LATENT DIMENSION COMPUTED AS IN [13] ON THE NONCAUSAL DATASET.

Completeness ↑						Informativness ↓			
model\gen	g_0	g_1	g_3	Avg.		g_0	g_1	g_2	Avg.
NC-ld-2	**0.94**	**0.84**	**1.00**	**0.92**		0.41	0.45	0.99	0.62
NC-ld-3	0.67	0.62	0.38	0.56		0.61	0.60	1.00	0.74
NC-ld-4	0.40	0.28	0.58	0.42		0.50	0.52	1.00	0.67
NC-ld-6	0.24	0.26	0.25	0.25		0.44	0.38	1.00	0.60
NC-ld-10	0.31	0.48	0.18	0.33		**0.40**	**0.38**	**0.98**	**0.59**

TABLE IV

COMPLETENESS (↑) AND INFORMATIVENESS (↓) SCORES [13] ON THE NONCAUSAL DATASET.

latent dimension encodes each generative factor. Using our measure we can not only calculate the usual disentanglement (e.g. which latent dimension corresponds to which generative factor) but can also precisely determine the coverage (e.g. how many classes of a generative factor are encoded in each dimension). We formalize this idea below.

The aim of this experiment was to determine if an SCM of a given dataset is preserved in a VAE's latent space. To this end, we labeled the datasets and trained a classifier on the decoded images. We then used *latent interventions* on a dimension in the VAE's latent space and the trained classifier to obtain the distribution over the generated labels (Figure 9 left). We compared this distribution to the distribution of labels obtained from the ground truth images where we fixed a class of a factor in its SCM (Figure 9 right). We call this a *class intervention*. Finally we calculate the similarity between these two distributions using Maximum Mean Discrepancy (MMD) [11].

The intuition behind this procedure is the following. If latent interventions on a dimension produced images with only limited class labels, we would conclude that it encodes a generative factor (which one can be determined from the ground truth data). On the other hand, if interventions on a latent dimension produced images with a uniform distribution over all classes, we would conclude that the latent dimension does not contribute to any generative factor. Using MMD we can not only determine which factor we intervened on (aka precision) but also which class of the generative factor this part of latent space encodes (aka recall). This information is summarised for every pair of latent dimension and generative factor in our novel *Interventional Score* measure. For example, in the Causal dataset, if a set of interventions on the dimension z_i matches all possible class interventions on X, the score for the pair $(0, \mathrm{X})$ will be 1. On the other hand, if it matches only one class intervention of X the score will be 0.5. If only half of interventions matches one class intervention of X, the score will be 0.3125. We present the details below.

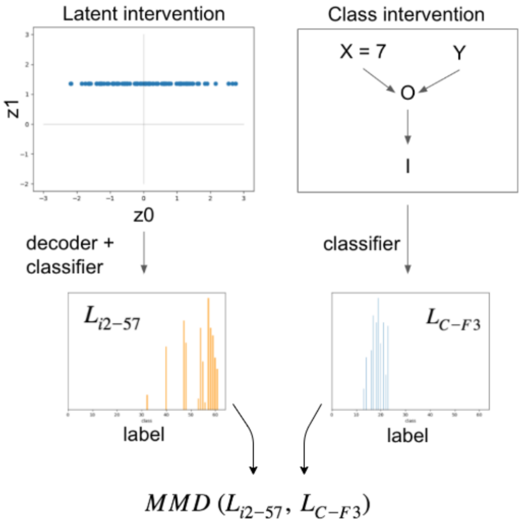

Fig. 9. Visualisation of class and latent interventions.

Classifier We labeled the images by the values of generative factors producing them. In the Causal dataset, where X and Y are categorically distributed on $[0, 31]$, we grouped these values into 8 equal bins. This yields $8 \cdot 8 = 64$ classes. In the NonCausal dataset we labeled X and Y in the same way but used 4 bins for the orientation factor O due to exploding combinatorial issue. Therefore, we obtained $8 \cdot 8 \cdot 4 = 256$ classes in the NonCausal dataset. We generated 10000 and 20000 images for the causal and noncausal cases, respectively, and processed them with each VAE. In this way we obtained 10 datasets $D_M = \{(M(I), y)\}$ for each of the 5 causal models $M = \textit{C-ld-n}$ and 5 noncausal models $M = \textit{NC-ld-n}$ where y denotes the label of the generated image I.

We trained a simple convolutional classifier C_M on each dataset D_M for 100 epochs with learning rate fixed to $1e-3$. The causal classifiers $C_{\textit{C-ld-n}}$ reached 100% accuracy, while this was not possible for the noncausal classifiers $C_{\textit{NC-ld-n}}$ due to poor VAE reconstructions. Since the accuracy of $C_{\textit{NC-ld-n}}$ on the validation set was below 50% for $n = 2, 3, 4, 6$, we carried out our interventional experiment only on $C_{\textit{NC-ld-10}}$ where it managed to achieve $100\%, 80\%$ accuracy on the training and validation sets, respectively.

Maximum Mean Discrepancy (MMD) [11] is a statistical test to determine whether or not two sets of samples were produced by the same distribution. The idea is to measure distances between distributions as distances between the mean values of all samples embedded in a feature space. More specifically, we map samples from both sets into another feature space called reproducing kernel Hilbert space, and compute the distance between mean values of the samples in each group. For the exact formula we refer the reader to Equation (3) in [11]. In our implementation we use a Gaussian kernel with the variance parameter $\alpha \in \{0.5, 0.75, 1, 1.5, 2\}$. Due to the nature of the Gaussian kernel, the lower the MMD score, the higher the similarity between the two sets (↓).

Interventional score is a novel measure of disentanglement as well as causal structure computed using latent interventions and MMD. In particular, assuming that a latent dimension z_i corresponds to a generative factor g_j, then an SCM is preserved in the latent space if an intervention on z_i produces only images from at most k different classes, where $k \in \{8, 24, 64\}$ depends on the intervened factor and the given SCM. This is because an intervention should fix $z_i = g_j$ to one class and therefore limit the variation of the generated images. In the Causal dataset k with always be equal to 8 as X and Y values are binned into 8 classes each. However, in the NonCausal dataset, intervening on X (or Y) can result in at most $8 \cdot 4 = 24$ classes while intervening on O can produce at most $8 \cdot 8 = 64$ classes. We encourage the reader to refer to Figure 9 for an easier understanding of the entire procedure described in the following.

We start by generating the ground truth datasets using class interventions. We intervene on each generative factor in the true SCM and set it to a value in each of the bins, e.g., we fix one class m for one factor F at a time. We then generate 200 images by uniformly sampling the rest of the generative factors, and obtain their classes from the corresponding classifier C_M. We denote the obtained set of labels by $L_{\textit{C-Fm}}$ and $L_{\textit{NC-Fm}}$ for the Causal and NonCausal SCMs, respectively. With slight abuse of notation we will simply write $L_{\textit{M-Fm}}$. In the causal case $M = C$, this yields $8 \cdot 8 = 64$ different sets of labels $\{L_{\textit{C-Fm}} : m = 1, \ldots, 8, F \in \{X, Y\}\}$, while in the noncausal case $M = NC$ we obtain $8 \cdot 8 \cdot 4 = 256$ sets $\{L_{\textit{NC-Fm}} : m = 1, \ldots, 8$ and $F \in \{X, Y\}$ or $m = 1, \ldots, 4$ and $F = O\}$ (Figure 9 right).

Next, we proceed with *latent interventions*. For a given VAE model $M \in \{\textit{C-ld-n}, \textit{NC-ld-n}\}$ for $n \in \{2, 3, 4, 6, 10\}$ we first sample 200 latent codes from the prior Gaussian distribution $N(0, 1)$. We then perform $t = 200$ interventions on each of the latent dimension z_i by setting $z_i = E$ where E is a value in an array $\mathscr{I} = \{-15 + e \cdot \frac{30}{t-1}\}_{e=0}^{t-1}$ of equidistant points from the interval $[-15, 15]$. We decode the latent codes and get their labels from the classifier C_M. The set of the obtained labels is denoted by L_{ip-e} where $p = 1, \ldots, n$ is a dimension in the latent space and e the position of the value E in the array of interventions $\mathscr{I}$. For a model M with n-dimensional latent space, this procedure yields $t \cdot n$ sets $\{L_{ip-e} : p = 1, \ldots, n$ and $e = 0, \ldots, t-1\}$ (Figure 9 left).

For each possible combination of two sets of labels (L_{ip-e}, L_{M-Fm}), we then compute the MMD scores. For each latent intervention *ip-e* we now wish to determine to which class intervention *Fm* the resulting images are closest to. This information is obtained by finding

$$\min_{Fm} MMD(L_{ip-e}, L_{M-Fm}),$$

which yields $p \cdot (t-1)$ pairs $(\textit{ip-e}, \textit{Fm})$.

Lastly, for each latent dimension p we wish to determine the number of times a latent intervention *ip-e* produced each generative factor. We can achieve this by decomposing the obtained pairs $(\textit{ip-e}, \textit{Fm})$ into a disjoint union

$$\bigsqcup_F \{(\textit{ip-e}, \textit{Fm})\} = \bigsqcup_F P_F.$$

For each latent dimension p we can now obtain the interventional statistics $(p, |P_F|, |P_F^{\neq m}|)$, where $P_F^{\neq m}$ is the subset of P_F containing only pairs with unique Fm elements, and $|S|$ denotes the cardinality of the set S. The *Interventional Score InS* is defined by computing the weighted average of the normalised statistics:

$$InS(p, F) = w \cdot \frac{|P_F|}{t} + (1 - w) \cdot \frac{|P_F^{\neq m}|)}{k} \tag{3}$$

where t is the number of interventions performed on the dimension p and k is the number of classes for the factor F. The weight w can be used to balance the importance of the "interventional precision" $\frac{|P_F|}{t}$ and "interventional recall" $\frac{|P_F^{\neq m}|)}{k}$.

We visualise the Interventional Score *InS* in the form of a Hinton diagram as shown in Figure 10 for the causal models *C-ld-n* and the noncausal *NC-ld-10* model. The reported score is the average *InS* over different $\alpha \in \{0.5, 0.75, 1, 1.5, 2\}$. The weight w was set to 0.5 and the larger the square the better (↑) the disentanglement and recall of the given pair (latent dimension, generative factor).

In Table V, we report the Interventional Score *InS*, interventional precision *InP*, interventional recall *InR* and the MMD score for *C-ld-4* obtained for $\alpha = 1$. We refer the reader to the Causal Girl's spreadsheet[3] for the results on the rest of the models.

[3] https://docs.google.com/spreadsheets/d/1J8GBr8jKBsHu0TmovNL_2fx25jClowSlVIgKM2vw8pI/edit?usp=sharing

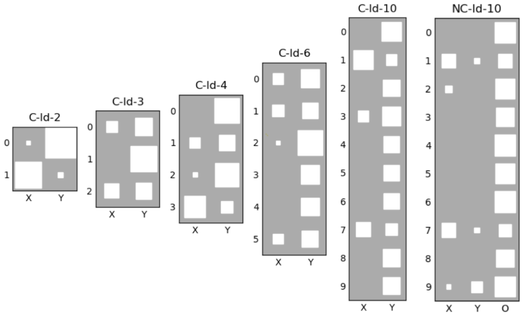

Fig. 10. Hinton diagram visualising Interventional Score (↑) obtained on the causal models *C-ld-n* as well as noncausal *NC-ld-10*.

(p, F)	*InS* ↑	*InP* ↑	*InR* ↑	*MMD* ↓
(0, X)	0.0	0.0	0.0	nan
(0, Y)	0.8125	1.0	0.625	0.209
(1, X)	0.3165	0.383	0.25	0.273
(1, Y)	0.496	0.617	0.375	0.17
(2, X)	0.167	0.209	0.125	0.687
(2, Y)	0.708	0.791	0.625	0.251
(3, X)	0.633	0.891	0.375	0.284
(3, X)	0.429	0.109	0.75	0.139

TABLE V

INTERVENTIONAL RESULTS ON *C-ld-4*. SEE FIGURE 10 FOR THE CORRESPONDING VISUALISATION OF *InS*.

D. *Finding Causal structures in the latent space*

We do a deeper investigation of the causal structure for the case of latent dimension equal to three. For this experiment we want to see if the third latent dimension in the causal case corresponds to the dependant factor O and compare the result to the non-causal case.

We identify which factors g_j are corresponding to z_i by fixing one of the generative factor, generating 1000 examples and analysing the variance of each latent dimension of their encodings. The idea is that if the representation is disentangled the variance in the dimension z_i that corresponds to the generative factor g_j should be smaller than in any other dimension. The results are shown in Table VI.

	Causal				Non-Causal		
data	z_0	z_1	z_2		z_0	z_1	z_2
X-constant	1.135	0.79	**0.569**		**0.231**	0.92	1.158
Y-constant	**0.475**	0.534	0.569		1.19	0.932	**0.268**
O-constant	-				1.219	**1.201**	1.279

TABLE VI

MAPPING GENERATIVE (X,Y,O) FACTORS TO LATENT DIMENSION (z_0,z_1,z_2) BY EVALUATING STD FOR EACH LATENT DIMENSION GIVEN DATA WITH ONE FACTOR HOLD CONSTANT.

Given this correspondence we train a simple single layer network in order to predict O given the values of X and Y. As we can see in SCM (1), O has a simple dependency on X and Y. Therefore the prediction should be straightforward to learn. We first predict the input O given X,Y for the ground truth causal and non-causal datasets in order to obtain a reference performance. Next, the same procedure is applied to predict the latent z_o given z_x and z_y. In this way we can investigate if the VAE mimics the causal graph using its latent representations.

Using the scipy [25] MLP regressor with 10 hidden nodes, ReLU activation, Adam optimizer and a learning rate of 0.01 was trained on 8500 known samples. The mean squared error on 1500 unseen test samples can be seen in Table VII.

	Causal	NonCausal
input predictions	3.4	1472.6
latent predictions	1163.3	1232.0

TABLE VII

THE MEAN SQUARED ERROR ON THE PREDICTION OF THE FACTOR O.

VI. DISCUSSION

As we could see in the Experiments, the Causality Girls are pretty good and should keep working on this.

A. *Disentanglement analysis*

In this section we discuss the disentanglement results obtained on the causal and noncausal models using measures from [12] and [13].

Distentanglement score [12] The best performing models according to this measure are *C-ld-2* and *NC-ld-10* for the causal and noncausal case, respectively. Surprisingly, we observe very poor performance of *NC-ld-3* which suggest that SCM is not respected in the latent space. We also point out that these results differ from the ones later obtained by using the measure of [13] which proves the difficulty of assessing the disentanglement of the representations.

Distentanglement score [13] We observe the same trend in causal (Table I) and noncausal models (Table III): the best score is achieved using *M-ld-2*, then the performance drops with increasing latent dimension until it somewhat recovers in *M-ld-10*. Looking at Hinton diagrams 7 and 8, we also see that 1) dimensions get ignored in the higher dimensional latent spaces, 2) the noncausal models seem to be disentangled a bit better, and 3) the clear trend with latent dimensions where *M-ld-10* somewhat recover. This suggest that the effect of the latent dimension to the disentanglement should be investigated in more detail. Moreover, since *NC-ld-3* performs worse than *NC-ld-2* we hypothesise that the orientation O is not considered as a generative factor in these models. This can be seen from Hinton diagrams as well as O seems to be treated in a similar way. One reason for such behaviour might be the poor quality of the noncausal VAEs as the obtained reconstructions are mostly circles or triangles as shown in Figure 4. Another possibility is that the orientation is a weaker factor compared to the position.

Completeness [13] We observe worse completeness score with increasing dimension which is as expected (Tables II and IV). However, we again observe a small recovery in *M-ld-10* models. It is also not clear why *NC-ld-2* again performs better than *NC-ld-3*, especially because in theory this should correspond to a failure case as the number of latent codes is smaller than the number of generative factors.

Informativeness [13] The informativeness score reported in Tables II and IV increases with increasing latent dimension which is the expected behaviour as more expressive models should perform better. We see that *M-ld-10* performs best, followed by *M-ld-2* models. In the causal case especially, the results are fairly random which makes it difficult to draw any conclusions.

Conclusion From the obtained results we conclude that the orientation does not seem to be recognised as a generative factor in the noncausal models. We point out two important questions:

- Are all generative factors equally important for a well performing generative process? If so, how can their importance be estimated beforehand?
- What is the role of the latent dimension and how does it influence the disentanglement?

B. Analysis of the causal structure

In this section we discuss whether or not the SCM was preserved in the latent space of the trained models.

Interventional score The most interesting observation from the Hinton diagrams visualising the obtained interventional score in Figure 10 is the dominance of the generative factor Y. We clearly see that X is underrepresented which becomes worse with increasing latent dimension. This is especially odd as the position factors X and Y are symmetric and they contribute equally to the generation of O. Note that this bias cannot be seen using the usual disentanglement measures by [22] and [13] as they do include the coverage information in their measures (as we do with the interventional recall). We leave the investigation of the observed bias for the future work.

Next, we see that the SCM in *C-ld-2* is indeed preserved (Figure 10): the interventions on the 0th and 1st latent dimensions correspond perfectly to Y and X, respectively. This conclusion is also supported with the previous results on the disentanglement.

Finally, we observe a bias towards the orientation factor O in *NC-ld-10*. This might be due to the class inbalance of the NonCausal dataset as O is binned into 4 classes only as opposed to 8 for X and Y. Thus a latent intervention is more likely to match a class intervention as the latter produces $8 \cdot 8 = 64$ different classes.

Causal search in *M-ld-3* Table VI shows that the variance in the latent dimension correctly identifies the corresponding generative factor as shown in Hinton diagrams in Figures 7 and 8. It also emphasises the difference between O, X and Y as O seems to be much harder to distinguish.

The results from Table VII further support the fact that that *M-ld-3* have not used the latent dimensions as in the underlying SCMs. In *C-ld-3*, the orientation factor O cannot anymore be predicted from X and Y in the latent space while this is an easy task in the input space. A potential reason for this is also poor disentanglement in the latent representations. Lastly, the results from the *NC-ld-3* might suggest that the model learned some spurious correlations but we emphasise that the absolute numbers in the classification error cannot be directly compared across different domains.

Conclusion We conclude that the underlying SCMs are violated in all models except for *C-ld-2*. This is somewhat expected as there is no mechanism in the VAE framework to accomodate such dependencies. We hypothesise that the results could be improved using a discriminator that would filter out bad reconstructions before inputting them to the classifier. In the current version, all reconstructions are used which can lead to meaningless labels for the invalid images.

VII. Potential Future Directions

In this section we identify future research directions and pose open question to the community.

A. Disentanglement measures

Measuring disentanglement is a difficult task which can be seen from the fact that there exists neither uniform definition nor evaluation framework adopted by the community. Most widely used definition of disentanglement is that one generative factor corresponds to exactly one latent code ([26], [15], [27], [28], [29], [30],[24] [31]) from which a number of different measurements emerged ([13], [32], [33], [12], [23]). In addition, these frameworks require that the generative factors are known a priori which real data is likely not to fulfill. Therefore, the community still investigates disentanglement by visually inspecting the latent traversals which are images obtained by changing one latent dimension at a time.

A potential reason why a unified disentanglement measure has not yet emerged could be precisely the restrictive assumptions. We identify two main problems:

1) The generative factors are not known for the given dataset. In this case the go to strategy is to disentangle as many features as possible by extracting *invariant features*. However, these do not necessarily correspond to underlying generative features. This approach is typical for supervised learning frameworks where disentangled representations are not considered to be essential for a good performance.

2) The generative factors are not independent of each other. The current definition where one generative factor maps to *exactly* one latent dimension necessarily requires generative factors to be *independent* from each other. Using SCMs, datasets can be described in a hierarchical way where only true causal effects are independent of each other while the rest of the variables are dependent. This can be a promising direction for a more flexible definition of disentanglement. However, not much work has been done to investigate if such causal structure can be captured by the latent variable models.

To determine if a causal structure could be imposed, we propose the following research questions:

- How to detect and impose the dependencies of the latent codes? The typical Gaussian prior works against this goal.

- Can the information given by a SCM be directly integrated into the VAE's ELBO loss?
- How would a causal latent space affect the generation of new samples? In ideal case, one could perform latent interventions to generate data with constraints.

B. Notion of Disentanglement

The accepted notion of disentanglement that stems from classical approaches [14] might not actually be perfectly suited for an unsupervised representation learning setting which is the core idea of the work in [32]. The authors combine the disentanglement with the ideas from the field of deep embedding which focuses on discovering representations that make a particular property explicit, often called *class*.

Let's demonstrate the inadequacy on a simple example of a dataset of different breeds of dogs. Deep embedding methods will try to find a representation that is based on the *class* (beagle, grayhound, Bernese Mountain Dog, etc) and to encode different breeds in different regions in the latent space. Such embedding can then be used to classify unseen examples of dogs by encoding them and finding the closest cluster of known examples. It is also often used to perform *few shot* learning, where a only a few examples of a class are introduced during the training. The core idea is that the embedding is able to extract the relevant features from the other classes such that they apply also to new unseen classes.

In the case of disentanglement, the goal is to separate the underling factors of generations without the usage of class labels. This means that in the dog breeds dataset the latent codes would represent distinct factors of variation of the breeds. For instance, one latent dimension could be responsible for the shape of the legs, one for the shape of the ears, another for the color of fur on the head. It would learn *building blocks* of what makes a dog independent of the breed it belongs to.

The assumption of independent generative factors now clearly becomes problematic. Dog breeds have certain constrains on what kind of *combinations* are acutely presented in the dataset. For example, longer legs seem to correlate with a slimmer body. If all generative factors of a dog were truly independent we would expect to have a lot of *Frankenstein-like*[4] dogs and a nearly uncountable number of different breeds. This example also shows that genes can be seen as generative factors whose exact combination for a desired outcome is unknown.

One can now see the trade off between the embedding and disentanglement approaches on problems where the separation of both classes and generative factors is desirable. Surprisingly not much work has been done investigating these trade-offs besides an older work [34] where *content* is separated from the *style*. Moreover, it is still not clear if an encoding where the latent codes share information about underlying factors is more valuable than a disentangled one. A starting point for such investigation could be the example presented in [32] where a factor $\Theta \in [0, 360]$ is encoded into two latent dimensions as $sin\Theta$ and $cos\Theta$ instead of a single dimension. This example is important also from the dimensionality perspective as many successful models have a much higher number of latent dimensions than (unknown) generative factors.

We propose the following concrete steps as potential research directions:

1) In order to validate if the combination of ideas from deep embedding and disentanglement fields are useful, one would need to design a dataset that is usable in both domains. Such a dataset would need to have known underlying generative factors as well as constrains that would assign a generated examples a unique class label. In this way, we could obtain a certain range of ”intra-class” variations. An inspiration could be taken from *procedural generated content* used in games [35].
2) Investigation if and how the causal structure could improve such a learning framework.
3) Analysis about if and when a disentangled representation is superior to a representation where generative factors are encoded into more than one latent dimensions, as well as investigating how the dimensionality of the latent space influences the disentanglement of the learned representations.

C. Integration of causal priors

Integrating causal priors into deep representation or deep embedding frameworks seems to be a underexplored research area. In [20] a Causal InfoGAN is introduced, however, the authors define causality as *sequential observations* and use the generative model to learn the probability of sequential observations $P_{data}(o, o\prime)$.

If our goal is to learn a latent space that captures the causal graph of the ground truth data, one could impose an additional loss matching the correlation between z_i and z_j to the corresponding correlation between g_i and g_j given a batch of samples B:

$$\mathscr{L}_{corr} = (G_{corr} - Z_{corr})^2 \tag{4}$$

$$G_{corr} = \left(\frac{\sum_{b=0}^{B} (g_i^b - \bar{g}_i)(g_j^b - \bar{g}_j)}{\sqrt{\sum_{b=0}^{B} (g_i^b - \bar{g}_i)^2 \sum_{b=0}^{B} (g_j^b - \bar{g}_j)^2}} \right)$$

$$Z_{corr} = \left(\frac{\sum_{b=0}^{B} (z_i^b - \bar{z}_i)(z_j^b - \bar{z}_j)}{\sqrt{\sum_{b=0}^{B} (g_i^b - \bar{z}_i)^2 \sum_{b=0}^{B} (z_j^b - \bar{z}_j)^2}} \right)$$

This approach is based of the assumptions that a mapping between g_j and z_i can be reliably established, the access to the ground truth SCM for the given dataset and a large enough batch size B such that the correlation are meaningful. Moreover, this would only be applicable to the case where the number of latent dimension is the same as the number of generative factors (or parentless nodes in the causal graph). Furthermore it needs to be validated if treating some latent dimensions as exogenous and others as endogenous variables is even possible.

[4] Technically we should refer to *Frankensteins Monster*

As an potential experiment we suggest to integrate this loss into the framework used in section V and compare if a better causal dependency can be achieved.

VIII. Conclusions

We investigated if the VAE Framwork preserves causal relationships present in the training data. We augmented the commonly used d-sprite dataset to a limited causal version where the rotation depends on the X-and Y-position. We evaluated the causal version as well as the non-causal dataset using three different disentanglement measures, two from the the literature as well as our own measure based on the Maximum Mean Discrepancy (MMD) between resulting distributions when intervention in the latent space. We also layed out potential future research directions.

References

[1] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[2] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking clever hans predictors and assessing what machines really learn," *Nature communications*, vol. 10, no. 1, pp. 1–8, 2019.

[3] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[4] M. Nissim, R. van Noord, and R. van der Goot, "Fair is better than sensational: Man is to doctor as woman is to doctor," *arXiv preprint arXiv:1905.09866*, 2019.

[5] Q. Zhang, Y. Nian Wu, and S.-C. Zhu, "Interpretable convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8827–8836.

[6] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv preprint arXiv:1708.08296*, 2017.

[7] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics surveys*, vol. 3, pp. 96–146, 2009.

[8] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *Int. Conf. Mach. Learn.*, 2014, pp. 1278–1286.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *Int. Conf. Learn. Represent.*, 2015.

[10] L. Matthey, I. Higgins, D. Hassabis, and A. Lerchner, "dsprites: Disentanglement testing sprites dataset," https://github.com/deepmind/dsprites-dataset/, 2017.

[11] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012. [Online]. Available: http://jmlr.org/papers/v13/gretton12a.html

[12] H. Kim and A. Mnih, "Disentangling by factorising," *arXiv preprint arXiv:1802.05983*, 2018.

[13] C. Eastwood and C. K. Williams, "A framework for the quantitative evaluation of disentangled representations," 2018.

[14] J. Schmidhuber, "Learning factorial codes by predictability minimization," *Neural Computation*, vol. 4, no. 6, pp. 863–879, 1992.

[15] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[16] M. D. Hoffman and M. J. Johnson, "Elbo surgery: yet another way to carve up the variational evidence lower bound," in *Workshop in Advances in Approximate Bayesian Inference, NIPS*, vol. 1, 2016, p. 2.

[17] M. Burns and J. Pearl, "Causal and diagnostic inferences: A comparison of validity," *Organizational Behavior and Human Performance*, vol. 28, no. 3, pp. 379–394, 1981.

[18] J. Pearl and M. Tarsi, "Structuring causal trees," *Journal of Complexity*, vol. 2, no. 1, pp. 60–77, 1986.

[19] D. Geiger and J. Pearl, "On the logic of causal models," in *Machine Intelligence and Pattern Recognition*. Elsevier, 1990, vol. 9, pp. 3–14.

[20] T. Kurutach, A. Tamar, G. Yang, S. J. Russell, and P. Abbeel, "Learning plannable representations with causal infogan," in *Advances in Neural Information Processing Systems*, 2018, pp. 8733–8744.

[21] J. Pearl, *Causality*. Cambridge university press, 2009.

[22] V. Thomas, J. Pondard, E. Bengio, M. Sarfati, P. Beaudoin, M.-J. Meurs, J. Pineau, D. Precup, and Y. Bengio, "Independently controllable factors," *arXiv preprint arXiv:1708.01289*, 2017.

[23] R. Suter, D. Miladinović, B. Schölkopf, and S. Bauer, "Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness," *arXiv preprint arXiv:1811.00007*, 2018.

[24] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework." *Iclr*, vol. 2, no. 5, p. 6, 2017.

[25] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. Jarrod Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. Carey, İ. Polat, Y. Feng, E. W. Moore, J. Vand erPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and S. . . Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[26] G. Desjardins, A. Courville, and Y. Bengio, "Disentangling factors of variation via generative entangling," *arXiv preprint arXiv:1210.5474*, 2012.

[27] S. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in *International Conference on Machine Learning*, 2014, pp. 1431–1439.

[28] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," *arXiv preprint arXiv:1512.09300*, 2015.

[29] J. Yang, S. E. Reed, M.-H. Yang, and H. Lee, "Weakly-supervised disentangling with recurrent transformations for 3d view synthesis," in *Advances in Neural Information Processing Systems*, 2015, pp. 1099–1107.

[30] W. F. Whitney, M. Chang, T. Kulkarni, and J. B. Tenenbaum, "Understanding visual concepts with continuation learning," *arXiv preprint arXiv:1602.06822*, 2016.

[31] E. L. Denton *et al.*, "Unsupervised learning of disentangled representations from video," in *Advances in neural information processing systems*, 2017, pp. 4414–4423.

[32] K. Ridgeway and M. C. Mozer, "Learning deep disentangled embeddings with the f-statistic loss," in *Advances in Neural Information Processing Systems*, 2018, pp. 185–194.

[33] T. Q. Chen, X. Li, R. B. Grosse, and D. K. Duvenaud, "Isolating sources of disentanglement in variational autoencoders," in *Advances in Neural Information Processing Systems*, 2018, pp. 2610–2620.

[34] J. B. Tenenbaum and W. T. Freeman, "Separating style and content with bilinear models," *Neural Computation*, vol. 12, no. 6, pp. 1247–1283, 2000.

[35] T. Short and T. Adams, *Procedural generation in game design*. CRC Press, 2017.