Causal Representation Learning: A Quick Survey

Frank Doehner

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
frank.doehner@kit.edu

Abstract

Causal representation learning (CRL) has recently become an object of intensive research. Representation learning aims to infer a lower dimensional, but meaningful, representation from a given set of data, effectively increasing interpretability and processability. Disentangled representation learning applies an independency constraint onto the inferred latent variables of the representation. The applicability of such frameworks on real world data is limited, as absolute independence between all generating factors is rarely the case. CRL assumes causal relations between these latent factors making it more flexible and suitable for real world settings. In this work we give an overview over several approaches to disentangled representation learning and give a short introduction to variational auto-encoders and generative adversarial networks. We follow up by covering the current state-of-the-art CRL frameworks and finish with remarks regarding current weaknesses of CRL as well as potential research topics.

1 Introduction

In recent years a myriad of machine learning approaches were developed with the goal of solving all kinds of different problem settings, including, among others, tasks of classification [25], regression [8] and clustering [23]. All of

these different machine learning tasks have a heavy performance dependency on the choice of data (or feature) representation in common. Feature engineering offers a way to incorporate prior expert knowledge into data, leveraging human ingenuity at the cost of exceedingly labor intensive work. Representation learning is a broad term comprising techniques that learn data representations, from which it is easier to extract valuable information. In a sense, obtaining a more useful representation can be understood as gaining a better understanding of the underlying physical or logical mechanisms that produce the data. For example an image of an arbitrary object such as an apple. A human does not need to see the entirety of an image, as i.e. the outline of the apple will be sufficient for classification. Other meaningful characteristics could be e.g. an object's colour and its surface characteristics. Therefore, useful feature representations are generally of lower dimension than the original data but provide higher level information. Studies have shown that disentangling of the feature representations leads to better generalization and performance in neural networks [10, 28]. This aligns with the intuition that complex data is given birth by the rich interactions of comparably few explaining factors. At the same time the interpretability of the neural networks is improved, as changes in the output can often be traced back to a single or a small number of explaining factors. The challenge and topic of much research lies in developing methods that infer these disentangled representations. Disentangled representation learning (DRL) has many benefits but it assumes independent underlying factors. While this assumption holds or offers a sufficiently close approximation for many tasks, it does not for many more complex real world settings. Instead, the explanatory factors are often causally related. In economics, for example inflation, interest rates, employment levels, and consumer spending are fundamentally independent factors but can drastically influence each other. Another example are e.g. shadows in an image that depend on the positions of light sources as well as the shapes and positions of the illuminated objects. All of these causal relations are generally ignored when learning a typical disentangled representation. By enforcing these additional, causal restraints on the neural networks during training, further performance improvements can be achieved. Despite major efforts in DRL, as well as in research on causal discovery, very few studies have been done on causal representation learning (CRL). The scope of this survey includes a brief overview of

neural network-based DRL techniques, a survey on studies tackling CRL and finally an outlook over possible future research on the topic.

2 Disentangled Representation Learning

Representation learning and especially DRL has been a research topic of much interest over the last ten years. Bengio et al. [2] defined a disentangled representation as a number of distinct, independent, informative and generative factors, which are invariant to change in other generating factors. A more mathematical definition based on group theory was later proposed by Higgins et al. [11]. Several different frameworks exist for inferring disentangled representations. The majority of those are based on either the variational auto-encoder (VAE), proposed by Kingam and Welling [14] or the generative adversarial network (GAN), proposed by Goodfellow et al. [9]. In the following we will give a brief overview over relevant VAE and GAN-based models. A comprehensive overview of the topic of DRL was published by Wang et al. [28].

2.1 Variational Auto-Encoder

VAEs combine the structure of an auto-encoder [17] with variational inference. A stochastic encoder learns the parameters ϕ of the variational posterior distribution in order to map the data distribution $\mathbf{x}$ to the latent representation $\mathbf{z}$. The generative decoder on the other hand attempts to recreate the input given the latent representation by learning the parameters θ of the joint distribution $p_\theta(\mathbf{x}, \mathbf{z}) = p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$, where $p_\theta(\mathbf{z})$ is the prior distribution over the latent space. The VAE attempts to maximize the log-likelihood of the data distribution $\log p_\theta(\mathbf{x})$ (2.1).

$$\log p_\theta(\mathbf{x}) = D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||(p_\theta(\mathbf{z}|\mathbf{x})) + \mathcal{L}_{\theta,\phi}(\mathbf{x}) \tag{2.1}$$

The by definition non negative Kulback-Leibler (KL) divergence D_{KL} pushes the variational posterior distribution $q_\phi(\mathbf{z}|\mathbf{x})$ to approximate the true posterior distribution $p_\theta(\mathbf{z}|\mathbf{x})$. $\mathcal{L}_{\theta,\phi}(\mathbf{x})$ is the evidence lower bound (ELBO). In practice

the ELBO (2.2) is maximized in order to generate a tight bound on the log-likelihood $\log p_\theta(\mathbf{x})$.

$$\mathcal{L}_{\theta,\phi}(\mathbf{x}) = -D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z})) + \mathbb{E}_{q_\phi(\mathbf{x}|\mathbf{z})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] \tag{2.2}$$

The first term in equation 2.2 penalises the distance between the variational $q_\phi(\mathbf{z}|\mathbf{x})$ and the prior distribution over the latent space $p_\theta(\mathbf{z})$. The second term on the other hand penalises the VAEs reconstruction error. Training such a VAE using stochastic gradient descent seems to be impossible at first due to the decoder's input $\mathbf{z}$ being sampled from the prior distribution over the latent space. The so called Reparameterization Trick by Kingma and Welling [14] bypasses this issue by multiplying the variance σ with a randomly sampled $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ instead of directly sampling $\mathbf{z}$.

While vanilla VAEs are generally able to infer disentangled representation, they rarely do without further constraints, especially when the data reaches a certain level of complexity. Higgins et al. [11] propose β-VAE, which introduces a penalty coefficient to the KL-divergence term in equation 2.2 which effectively encourages disentangled latent variables, but at the same time leads to a greater reconstruction error. Burgess et al. [4] approach the design of the loss function by simultaniously optimizing the mutual information between the input and task objective, as well as the mutual between the input and the latent space. In practice they subtract a linear parameter from the KL-divergence term in equation 2.2 and increases its value during network training. This leads to an improvement in the networks's reconstruction ability while still enforcing satisfactory disentanglement of the latent space. Kumar et al. [18] proposed DIP-VAE which, leverages an additional regularizer that penalizes the weighted distance between the marginal distribution of the variational posterior $q_\phi(\mathbf{z})$ and the latent prior $p_\theta(\mathbf{z})$. Kim et al. [13] introduced FactorVAE which utilizes the Total Correlation $D_{KL}(q_\phi(\mathbf{z})||\prod_j q_\phi(z_j))$, a measure of dimensional Independence, as an additional regularizer in the loss function. Chen et al. [5] propose β-TCVEA, in which they decompose the KL-divergence in equation 2.2 into three separately weighted terms as shown in equation 2.3.

$$\begin{aligned}\mathcal{L}_{\theta,\phi}(\mathbf{x}) = \mathbb{E}_{q_\phi(\mathbf{x}|\mathbf{z}),p_\theta(\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - \alpha I_q(\mathbf{z};\mathbf{x}) \\ - \beta D_{KL}(q_\phi(\mathbf{z})||\prod_j q_\phi(z_j)) - \gamma \sum_j D_{KL}(q_\phi(z_j)||p_\theta(z_j))\end{aligned} \tag{2.3}$$

α, β and γ are weights. The second term in equation 2.3 is the mutual information, the second term the Total Correlation and the third term a dimension wise KL-divergence.

2.2 Generative Adversarial Network

Another neural network architecture which researchers have adapted in order to infer disentangled representations is the GAN [9]. A GAN consists of two, with each other competing, neural networks. The generator network (G) creates data by sampling from a latent representation, while the discriminator network (D) attempts to differentiate real and generated data. As a result the two networks compete with each other in a zero-sum game. The training objective lies in the generator generating data which are indiscernible for the discriminator. Equation 2.4 shows a GAN's optimization objective.

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim P_{Data}}\left[\log D(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(1 - D(G(\mathbf{z}))\right)\right] \tag{2.4}$$

P_{Data} is the real dataset and $p(\mathbf{z})$ is the prior distribution of the latent space. Chen et al. [6] proposed infoGAN, one of the first GAN-based approaches to DRL. The authors differentiate between a noise variable $\mathbf{z}$ and a structured latent variable $\mathbf{c}$. They then add a weighted regularizing term $-\lambda I(c; G(c,z))$ that encourages the mutual information between $\mathbf{c}$ and the generated data to stay large. Jeon et al. [12] introduce IB-GAN which compresses the inferred latent representation by limiting the mutual information, effectively achieving better disentangled representations. Lin et al. [19] propose a self supervised model InfoGAN-CR which uses a contrastive regularizer. The model is trained by evaluating generated images for which one generating latent factor is kept constant while all the others are randomly sampled. The contrastive regularizer then discerns the constant latent variable effectively encouraging the latent variables to be meaningful and distinct. Zhu et al. [34] proposes PS-SC GAN which utilizes spacial constrictions in the form of masks in order to localize the effect of latent variables in an image. Furthermore, in order for the disentangled latent space to encode simple and distinct variations in the data, they employ

perceptual simplicity by imposing small perturbations on single latent variables and identify them in the generated images. Finally Wei et al. [29] build a regularizer with a Jacobian of the generators output with respect to the latent input. The regularizer measures orthogonality of the Jacobian vectors, effectively rewarding independence of the latent dimensions.

3 Causal Representation Learning

Locatello et al. [21] argue that disentangled representations can only be inferred under inductive biases, both on the learning approach and on the data. They further query the mutual independence necessity of latent variables in DRL. CRL casts aside this independence assumption of the latent generating factors. Instead, the latent variables are dependent on each other, in accordance to an underlying causal mechanism. Shen et al. [26] proved that models with an independent latent prior distribution are not identifiable. CRL approaches condition their latent prior distributions by enforcing causal structure in different ways. Träuble et al. [27] further show that most DRL approaches are unable to disentangle latent factors if correlation exists in the data.

Yang et al. [30] were first to propose a representation learning framework which learns a structural causal model (SCM) [24] under light supervision. Their CausalVAE framework includes a SCM layer, which takes the independent exogenous factors $\boldsymbol{\epsilon}$ from the VAE decoder and transforms them into structured causal representations $\mathbf{z}$ which follow the structure of a directed acyclic graph (DAG). This DAG structure can be formulated as a strictly upper triangular adjacency matrix $\mathbf{A}$. The mathematical formulation of the SCM layer is shown in equation 3.1.

$$\mathbf{z} = \mathbf{A}^T\mathbf{z} + \boldsymbol{\epsilon} = (\mathbf{I} - \mathbf{A}^T)^{-1}\boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}) \tag{3.1}$$

$\mathbf{A}$ is the to be learned adjacency matrix and $\mathbf{I}$ the identity Matrix. In a next step, prior to being passed to the decoder, the causal representation $\mathbf{z}$ is passed to a Mask Layer where it reconstructs itself starting from the independent exogenous variables and following up with the on their parents dependent endogenous

variables according to the DAG structure. This process is described by equation 3.2.

$$z_i = g_i(\mathbf{A}_i \circ \mathbf{z}) + \epsilon_i \tag{3.2}$$

$\circ$ is the elementwise multiplication between the weight vectors $\mathbf{A}_i$ and $\mathbf{z}$ with $\mathbf{A} = [\mathbf{A}_1|...|\mathbf{A}_n]$ and g_i is a set of mildly nonlinear functions. The Mask Layer, besides enforcing the DAG characteristics on the causal representation $\mathbf{z}$, allows for interventions to be performed. In causality an intervention refers to the act of manipulating a variable in the system by external means. The outcomes of such interventions, besides being of interest themselves, can offer much insight on the underlying causal structure of a system. The weak supervision is realized through the labels u_i which hold the true causal concepts. They are fed to the encoder together with the data $\boldsymbol{\epsilon} = h(\mathbf{z}, \mathbf{u}) + \boldsymbol{\zeta}$ and further act as a constraint on the latent prior distribution $p_\theta(\mathbf{z}|\mathbf{u})$. Besides the ELBO objective (3.3) Yang et al. introduce three other regularization terms (3.4).

$$\begin{aligned} \textbf{ELBO} =& \mathbb{E}_{q\chi}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \\ & D_{KL}(q_\phi(\boldsymbol{\epsilon}|\mathbf{x},\mathbf{u})||p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})) \\ & D_{KL}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})||p_\theta(\mathbf{z}|\mathbf{u}))] \end{aligned} \tag{3.3}$$

$$\begin{aligned} \mathcal{L} = & - \textbf{ELBO} + \alpha \left(tr((\mathbf{I} + \frac{c}{m}\mathbf{A} \circ \mathbf{A})^n) - n \right) \\ & + \beta \mathbb{E}_{q\chi} \|\mathbf{u} - \sigma(\mathbf{A}^T\mathbf{u})\|_2^2 \\ & + \gamma \mathbb{E}_{\mathbf{z}\sim q_\phi} \sum_{i=1}^{n} \|z_i - g_i(\mathbf{A}_i \circ \mathbf{z})\|^2 \end{aligned} \tag{3.4}$$

α, β and γ are hyperparameters. The second term of equation 3.4 becomes zero iff the adjacency matrix $\mathbf{A}$ corresponds to a DAG. Zheng et al. [33] were the first to propose a continuous formulation of the DAG constraint for marticies. Yu et al. [31] then further developed the this formulation into the the form as shown in equation 3.4. The third and fourth term ease the training task of the two unknown variables $\mathbf{A}$ and $\mathbf{z}$ by encouraging $\mathbf{A}$ to correctly describe the causal relations of the labels $\mathbf{u}$, where σ is the logistic function and χ the data set, and $\mathbf{z}$ to be accurately reproduced by the Mask Layer. In addition to the CausalVAE

framework the authors provide a set of conditions under which their model is identifiable.

Komanduri et al. [16] propose the SCM-VAE framework. Contrary to the CausalVAE, SCM-VAE assumes that the underlying SCM is given and therefore does not need to be learnt during network training. Instead, it is utilized as a constrained on the latent prior distribution $p_\theta(\mathbf{z}|\mathbf{u})$. Similar to CausalVAE SCM-VAE incorporates a Mask Layer into the decoder, which allows for interventions to be performed on the SCM. Furthermore SCM-VAE is not limited by a linearity constraint on the SCM.

Recently, Komanduri et al. [15] improved on their SCM-VAE framework in the form of ICM-VAE. While they fundamentally assume the same given information, labels of the causal variables and the corresponding SCM, ICM-VAE improves on its predecessor by implementing independent causal mechanisms through learnable flow-based diffeomorphic functions. This implementation of structural causal flow allows for more complex models than the strictly additive noise model, which CausalVAE and SCM-VAE follow.

Similarly Fan et al. [7] propose another flow-based framework called Cauf-VAE, which does not require the underlying SCM as prior information, but infers it during training.

Brehmer and Haan et al. [3] propose an intervention-based framework for CRL. Instead of labels for the generating latent variables or the causal structure itself, they leverage datasets with paired samples from before and after random, unknown interventions. The authors define latent causal models (LCMs) as sets of an acyclic SCM $\mathcal{C}$, an observation space $\mathcal{X}$, a diffeomorphic decoder g, a set of interventions $\mathcal{I}$ on $\mathcal{C}$ and a probability measure $p_\mathcal{I}$ over $\mathcal{I}$. They then show, that under some mild constraints over the LCMs, two LCMs $\mathcal{M}$ and $\mathcal{M}'$ are equal up to a relabeling and an elementwise transformation of the causal variables if the LCMs entail equal weakly supervised distributions $p^\mathcal{X}_\mathcal{M}(\mathbf{x}, \tilde{\mathbf{x}}) = p^\mathcal{X}_{\mathcal{M}'}(\mathbf{x}, \tilde{\mathbf{x}})$ and vice versa. $\mathbf{x}$ and $\tilde{\mathbf{x}}$ are the observations before and after an intervention. Beyond a theoretical prove that a system's causal structure can be learnt under their weakly supervised setting, the authors introduce implicit latent causal models (ILCMs). Similar to the chicken-and-egg problem - which came first? - it is difficult to jointly learn the causal variables and the causal structure of a system

in an explicit fashion. The ILCM implicitly represents the causal structure through a neural solution function $s(\mathbf{e}) = \mathbf{z}$ which maps the vector of exogenous noise variables to the endogenous causal variables. The latent variables in an ILCM are therefore noise encodings defined by the inverse solution function. By learning the transformation $\tilde{\mathbf{z}}_i = \overline{s}(\tilde{\mathbf{e}}_i; \mathbf{e})$ in the ILCM, the solution function of a corresponding, unique explicit LCM (ELCM) can be recovered. the causal graph can be extracted from the ILCM after training by either using intervention-based causal discovery algorithms on the learned representations or by iterative topological ordering of the solution functions s_i.

An et al. [1] show, that in order to achieve causally sound generation through a decoder, a disentangled latent space is insufficient. Due to the entangled decoder, it is not given that the generated data indeed follows the inferred causal structure of the latent space. For instance, a do-operation on a single variable in latent space might affect that variable's parents, as the natural decoder has no regularization preventing this from happening. Their framework CDG-VAE allows for correct causal generation under the rather strict assumption of comprehensive knowledge over the causal information for supervision.

Lippe et al. [20] tackle the problem of CRL from temporal sequences. Previous works have made the naive assumption that intervening on a variable only has an effect in later time steps. This is often inaccurate, as there are often causal effects in real world settings which act faster than the modeled time steps. The authors show that by using interventions, effectively removing instantaneous effects of parent variables, they can recover the minimal causal variables and identify the caual graph under mild assumptions.

Louizos et al. [22] and Zhang et al. [32] provide representation learning frameworks for treatment effect estimation. Treatment effect estimation tries to predict the effect of a certain treatment under hidden confounding. Both works assume a surrogate rich setting, ample indirect information about the hidden confounding variable. While they are technically not learning a causal representation, their models CEVAE [22] and TEDVAE [32] hold the causal structure of the setting implicitly, and learn the latent hidden confounder during training. TED-VAE improves upon CEVAE by further distinguishing between multiple latent variables depending on their effect on either/and treatment and outcome.

While the field of CRL is mainly dominated by VAE-based approaches Shen et al. [26] introduced DEAR, a GAN-based CRL framework. They employ labels of the causal variables for supervision and assume a super-graph of the underlying causal graph is known. The DEAR's loss function is given in equation 3.5.

$$\min_{E,G,F} L_{E,G,F} := D_{KL}(q_E(\mathbf{x},\mathbf{z}), p_{G,F}(\mathbf{x},\mathbf{z})) + \lambda \mathbb{E}_{\mathbf{x},\mathbf{y}}[l_s(E;\mathbf{x},\mathbf{y})] \tag{3.5}$$

E references the encoder network, G the generative network and F the causal prior. The encoded joint distribution factorizes as $q_E(\mathbf{x},\mathbf{z}) = q_{\mathbf{x}}(\mathbf{x})q_E(\mathbf{z}|\mathbf{x})$ and the generated joint distribution as $p_{G,F}(\mathbf{x},\mathbf{z}) = p_F(\mathbf{z})p_G(\mathbf{x}|\mathbf{z})$, where

$$p_F(\mathbf{z}) = f((\mathbf{I} - \mathbf{A}^T)^{-1}h(\epsilon))$$

encodes the generally nonlinear SCM. f and h are generally nonlinear element-wise transformations, $\mathbf{A}$ is the weighted adjacency matrix corresponding to the causal graph and $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$. The second term of equation 3.5 is a supervised regularizer. λ is a hyperparameter, $\mathbf{y}$ are the supervision labels and l_s a function which penalizes deviation from the labels. In the case of binary labels the authors chose a cross entropy loss. Like many of the other CRL frameworks, DEAR allows for interventions to be performed on the generating latent variables.

4 Possible Research Ideas and Final Remarks

Even though CRL is still a rather young research topic it has obtained much interest from the scientific community. Compared to general DRL, CRL is much more applicable to real world problems, as causal relations between generating latent factors are much more frequent than the contrary. On the other hand, current CRL frameworks are limited by their need for supervision in order to guarantee identifiability. Labels for all latent causal variables or for example the complete causal graph are rather strict requirements which cannot always be fulfilled in real world settings. Similarly, datasets of paired samples from before and after interventions might be feasible for some of the generating variables but rarely for all. Therefore, a framework which allows for not only a single type

of supervision but a mixture of different types of supervision would make CRL frameworks much more applicable. Another scenario which, to the best of our knowledge, has not been covered yet, is the case of ample supervision on all but one generating variable. Current approaches require some form of supervision for all generating variables. In real world problems, such as manufacturing processes, this cannot be guaranteed as there might be further, hidden variables which have not been considered. In future research we plan to investigate CRL-based methods for inferring quantitative or even qualitative information on such hidden variables. Attaining knowledge about the existence or even better, about the placement of such an additional hidden variable within the causal graph would be very valuable. Once discovered, additional measurements or even an instrumentation could be done in order to obtain additional supervision labels.

References

[1] SeungHwan An, Kyungwoo Song, and Jong-June Jeon. "Causally Disentangled Generative Variational AutoEncoder". In: *ArXiv* abs/2302.11737 (2023). URL: https://api.semanticscholar.org/CorpusID:257102874.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation Learning: A Review and New Perspectives". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.8 (2013), pp. 1798–1828. DOI: 10.1109/TPAMI.2013.50.

[3] Johann Brehmer et al. "Weakly supervised causal representation learning". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 38319–38331. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/fa567e2b2c870f8f09a87b6e73370869-Paper-Conference.pdf.

[4] Christopher P Burgess et al. "Understanding disentangling in β-VAE". In: *arXiv preprint arXiv:1804.03599* (2018).

[5] Ricky TQ Chen et al. “Isolating sources of disentanglement in variational autoencoders”. In: *Advances in neural information processing systems* 31 (2018).

[6] Xi Chen et al. “Infogan: Interpretable representation learning by information maximizing generative adversarial nets”. In: *Advances in neural information processing systems* 29 (2016).

[7] Di Fan, Yannian Hou, and Chuanhou Gao. “CF-VAE: Causal Disentangled Representation Learning with VAE and Causal Flows”. In: *arXiv preprint arXiv:2304.09010* (2023).

[8] Manuel Fernández-Delgado et al. “An extensive experimental survey of regression methods”. In: *Neural Networks* 111 (2019), pp. 11–34.

[9] Ian Goodfellow et al. “Generative adversarial nets”. In: *Advances in neural information processing systems* 27 (2014).

[10] Irina Higgins et al. “beta-vae: Learning basic visual concepts with a constrained variational framework”. In: *International conference on learning representations*. 2016.

[11] Irina Higgins et al. “Towards a Definition of Disentangled Representations”. In: *CoRR* abs/1812.02230 (2018). arXiv: 1812.02230. URL: http://arxiv.org/abs/1812.02230.

[12] Insu Jeon et al. “Ib-gan: Disentangled representation learning with information bottleneck generative adversarial networks”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 9. 2021, pp. 7926–7934.

[13] Hyunjik Kim and Andriy Mnih. “Disentangling by factorising”. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2649–2658.

[14] Diederik P Kingma and Max Welling. “Auto-encoding variational bayes”. In: *arXiv preprint arXiv:1312.6114* (2013).

[15] Aneesh Komanduri et al. “Learning Causally Disentangled Representations via the Principle of Independent Causal Mechanisms”. In: *arXiv preprint arXiv:2306.01213* (2023).

[16] Aneesh Komanduri et al. “SCM-VAE: Learning Identifiable Causal Representations via Structural Knowledge”. In: *2022 IEEE International Conference on Big Data (Big Data)*. IEEE. 2022, pp. 1014–1023.

[17] Mark A Kramer. “Nonlinear principal component analysis using autoassociative neural networks”. In: *AIChE journal* 37.2 (1991), pp. 233–243.

[18] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. “Variational inference of disentangled latent concepts from unlabeled observations”. In: *arXiv preprint arXiv:1711.00848* (2017).

[19] Zinan Lin et al. “Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans”. In: *international conference on machine learning*. PMLR. 2020, pp. 6127–6139.

[20] Phillip Lippe et al. “Causal representation learning for instantaneous and temporal effects in interactive systems”. In: *The Eleventh International Conference on Learning Representations*. 2022.

[21] Francesco Locatello et al. “Challenging common assumptions in the unsupervised learning of disentangled representations”. In: *international conference on machine learning*. PMLR. 2019, pp. 4114–4124.

[22] Christos Louizos et al. “Causal effect inference with deep latent-variable models”. In: *Advances in neural information processing systems* 30 (2017).

[23] Erxue Min et al. “A survey of clustering with deep learning: From the perspective of network architecture”. In: *IEEE Access* 6 (2018), pp. 39501–39514.

[24] Judea Pearl. *Causality*. Cambridge university press, 2009.

[25] Waseem Rawat and Zenghui Wang. “Deep convolutional neural networks for image classification: A comprehensive review”. In: *Neural computation* 29.9 (2017), pp. 2352–2449.

[26] Xinwei Shen et al. “Weakly supervised disentangled generative causal representation learning”. In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 10994–11048.

[27] Frederik Träuble et al. “On disentangled representations learned from correlated data”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 10401–10412.

[28] Xin Wang et al. “Disentangled representation learning”. In: *arXiv preprint arXiv:2211.11695* (2022).

[29] Yuxiang Wei et al. “Orthogonal jacobian regularization for unsupervised disentanglement in image generation”. In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 6721–6730.

[30] Mengyue Yang et al. “Causalvae: Disentangled representation learning via neural structural causal models”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 9593–9602.

[31] Yue Yu et al. “DAG-GNN: DAG structure learning with graph neural networks”. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 7154–7163.

[32] Weijia Zhang, Lin Liu, and Jiuyong Li. “Treatment effect estimation with disentangled latent factors”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12. 2021, pp. 10923–10930.

[33] Xun Zheng et al. “Dags with no tears: Continuous optimization for structure learning”. In: *Advances in neural information processing systems* 31 (2018).

[34] Xinqi Zhu, Chang Xu, and Dacheng Tao. “Where and what? examining interpretable disentangled representations”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5861–5870.