Do learned representations respect causal relationships?

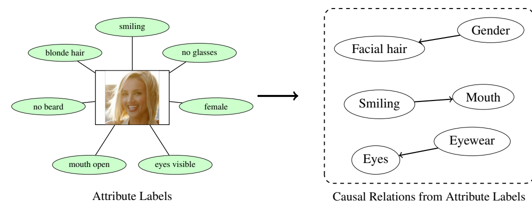
Lan Wang and Vishnu Naresh Boddeti
Michigan State University
wanglan3,vishnu@msu.edu

Abstract

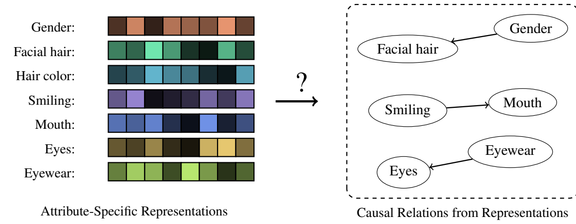
Data often has many semantic attributes that are causally associated with each other. But do attribute-specific learned representations of data also respect the same causal relations? We answer this question in three steps. First, we introduce NCINet, an approach for observational causal discovery from high-dimensional data. It is trained purely on synthetically generated representations and can be applied to real representations, and is specifically designed to mitigate the domain gap between the two. Second, we apply NCINet to identify the causal relations between image representations of different pairs of attributes with known and unknown causal relations between the labels. For this purpose, we consider image representations learned for predicting attributes on the 3D Shapes, CelebA, and the CASIA-WebFace datasets, which we annotate with multiple multi-class attributes. Third, we analyze the effect on the underlying causal relation between learned representations induced by various design choices in representation learning. Our experiments indicate that (1) NCINet significantly outperforms existing observational causal discovery approaches for estimating the causal relation between pairs of random samples, both in the presence and absence of an unobserved confounder, (2) under controlled scenarios, learned representations can indeed satisfy the underlying causal relations between their respective labels, and (3) the causal relations are positively correlated with the predictive capability of the representations. Code and annotations are available at: https://github.com/human-analysis/causal-relations-between-representations.

1. Introduction

Consider the face image in Fig. 1a. Automated face analysis systems typically involve extracting semantic attributes from the face. These attributes are often related through an underlying causal mechanism governing the relations between them. Modern computer vision systems excel at predicting such attributes by learning from large-scale



(a) Illustration of causal relations between attribute labels



(b) Causal relations between attribute-specific learned representations

Figure 1. Visual data may have multiple causally associated attributes. The goal of this paper is to determine whether attribute-specific learned representations respect the underlying causal relationships between the attributes? And if so, to what extent?

annotated datasets. This is achieved by learning compact attribute-specific representations of the image from which the attribute prediction is made. This setting naturally raises the following questions (Fig. 1b): (1) *Can we estimate the causal relations between high-dimensional representations purely from observational data with high accuracy?*, (2) *Do the learned attribute-specific representations also satisfy the same underlying causal relations, and if so to what extent?*, and finally (3) *How are the causal relations affected by factors such as the extent of training, overfitting, network architecture, etc.* Answering these questions is the primary goal of this paper.

Our work is motivated by the empirical observation that modern representation learning algorithms are inclined to uncontrollably absorb all correlations in the data [50]. Consequently, while such systems have exhibited significantly improved empirical performance across many applications, it has also led to unintended consequences, ranging from bias against demographic groups [4] to loss of privacy by extracting and leaking sensitive information [49]. Identifying the causal relations between representations can help mitigate the deleterious effects of spurious correlations. With the proliferation of computer vision systems that employ such representations, it is imperative to devise tools to discover the causal relations given a set of representations.

Discovering causal relations from learned representations poses two main challenges. First, causal discovery typically involves interventions [44] on the data which are either difficult or impossible on the observational representation space. For instance, in the image space, interventions may be possible during the image acquisition process for certain attributes such as hair color, eyeglasses, etc. Such interventions are, perhaps, not possible for attributes such as gender or ethnicity. On the other hand, it is not apparent how to intervene on any of these attributes directly in the representation space. Second, causal discovery methods, for pairs or whole graphs, are typically evaluated on small-scale low-dimensional datasets with multiple related attributes. However, there are no large-scale image datasets labeled with multiple causally associated attributes, nor are there any standardized protocols for evaluating the effectiveness of causal discovery methods on learned representations. While existing datasets such as MSCOCO [32] and CelebA [34] are labeled with multiple attributes they are either not causally related to each other (e.g., MSCOCO) or only have binary labels that suffer from severe class imbalance (e.g., CelebA).

To mitigate these challenges; (1) We propose Neural Causal Inference Net (NCINet) – a learning-based approach for observational causal discovery from high-dimensional representations, both in the presence and absence of a confounder. NCINet is trained on a custom synthetic dataset of representations generated through a known causal mechanism. And, to ensure that it generalizes to real representations with complex causal relations we, (a) incorporate a diverse set of function classes with varying complexity into the data generating mechanism, and (b) introduce a learning objective that is explicitly designed to encourage domain generalization. (2) We develop an experimental protocol where, (a) existing datasets can be controllably resampled to induce a desired known causal relationship between the attribute labels, (b) learn attribute representations from the resampled data and infer the causal relations between them. We adopt three image datasets, namely, 3D Shapes dataset [7], CelebA [34] and CASIA WebFace [60], where we annotate the latter with multiple multi-label attributes.

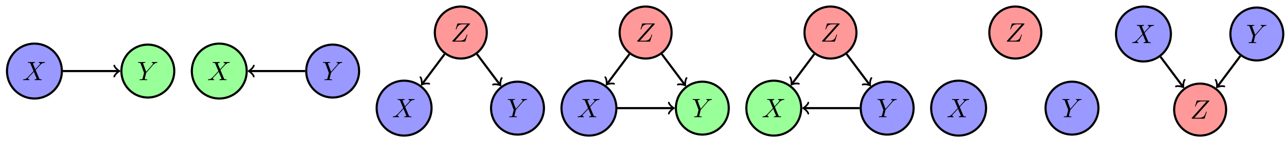
Contributions: First, we propose a learning-based tool, NCINet, for causal discovery from high-dimensional observational data, both in the presence and absence of a confounder. Numerical experiments on both synthetic and real-world data causal discovery problems indicate that NCINet exhibits significantly better causal inference generalization than existing approaches. Second, we employ NCINet for causal inference on attribute-specific learned representations and make the following observations; (1) Learned attribute-specific representations *do* satisfy the same causal relations between the corresponding attribute labels under controlled scenarios with high causal strength. (2) The causal consistency is highly correlated with the predictive capability of the attribute classifiers (e.g., causal consistency degrades with overfitting).

2. Related Work

Representation Learning: The quest to develop image representations that are simultaneously robust and discriminative has led to extensive research on this topic. Amongst the earliest learning-based approaches, Turk and Pentland proposed Eigenfaces [55] that relied on principal component analysis (PCA) of data. Later on, integrated and high-dimensional spatially local features became prevalent for image recognition, notable examples include local binary patterns (LBP) [1], scale-invariant feature transform (SIFT) [39] and histogram of oriented gradients (HoG) [12]. In contrast to these hand-designed representations, the past decade has witnessed the development of end-to-end representation learning systems. Representations learned from supervised learning [18, 33, 52], disentangled learning [8, 10, 19, 28, 54], and most recently self-supervised learning [9, 14, 16, 42, 43] now typify modern image representations. The goal of these approaches is to learn universal representations that generalize well across arbitrary tasks. Hence, they are inclined to uncontrollably learn all contextual correlations in data. Our goal in this paper is to verify whether learned representations retain the underlying causal relations of the data generating process.

Causal Inference: Randomized controlled experiments are the gold standard of causal inference. However, in many computer vision applications, we cannot control the image formation process rendering such experiments infeasible. Concurrently, a plethora of approaches have been proposed for causal discovery purely from observational data under two main settings, full graph or pairwise. Estimating the full causal graph has been thoroughly studied, both using learning-based [2, 3, 6, 24, 27] and non-learning-based [11, 26, 45, 51] approaches.

In this paper, we restrict our focus to causal discovery for the particular case where we have access to only two random variables at a time, both in the presence or absence of an unobserved confounder. Significant efforts have also been devoted to this problem under different scenarios. These include, comparison of information entropy for discrete variables [29, 30], neural causal methods [17, 36, 38], comparison of noise statistics in causal and anti-causal directions [20, 41], comparison of regression errors between the causal and anti-causal directions [4], comparison of Kolmogorov complexity [5, 56], building classi-



(a) $\mathcal{G}_1$, Label: 1 (b) $\mathcal{G}_2$, Label: 2 (c) $\mathcal{G}_3$, Label: 0 (d) $\mathcal{G}_4$, Label: 1 (e) $\mathcal{G}_5$, Label: 2 (f) $\mathcal{G}_6$, Label: 0 (g) $\mathcal{G}_7$, Label: 0

Figure 2. All possible causal relations between pairs of random variables. A node in blue denotes cause, green denotes effect, and red denotes confounder or common effect (Z). We only consider scenarios where we observe X and Y, but not Z. As such, the graphs represent three different causal relations, (i) Label 1: causal relation ($X \to Y$); (ii) Label 2: anti-causal relation ($X \leftarrow Y$); (iii) Label 0: X and Y are unassociated. Note that since Z is not observed, $\mathcal{G}_7$ is equivalent to $\mathcal{G}_6$ and can thus be ignored.

fication and regression trees [40], analyzing conditional distributions [15, 25] and many more [21, 35, 47]. A majority of the aforementioned methods have been designed and applied to low-dimensional variables, except for [5,21,40,56].

Within the broader context of computer vision, there is growing interest in causal discovery [36], causal data generation [31], incorporating causal concepts within scene understanding systems [53,57,59,62], domain adaptation [61], and debiasing [58].

In this paper, the proposed NCINet is a neural causal inference method that is tailored for high-dimensional variables. It incorporates, (1) direct supervision through causal labels, indirect supervision by comparing regression errors in the causal and anti-causal direction and an adversarial loss to encourage domain generalization, and (2) in contrast to all existing approaches, our model is trained to infer causal relations from all possible (see Fig. 2) pairwise cases, including in the presence and absence of an unobserved confounder.

3. Causal Relations Between Representations

First, we define the primary causal inference query that this paper seeks to answer i.e., *"Do learned representations respect causal relationships?"*. Consider the graph $\mathcal{G}_1$ in Fig. 2, which has two attributes X and Y, where the causal relation between them is $X \to Y$. An image $\boldsymbol{I}$ is generated by an unknown stochastic function of these two attributes. Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be high-dimensional attribute-specific representations learned for predicting labels X and Y, respectively, from the corresponding images. The structural causal equations (SCEs)[1] that characterize this process are:

$$\begin{aligned} a_x \sim P_c(X) \quad a_y \sim P_e(Y|X = a_x) \\ \boldsymbol{I} = g(a_x, a_y, \epsilon) \\ \boldsymbol{x} = h_X(\boldsymbol{I}; \boldsymbol{\theta}_X) \quad \boldsymbol{y} = h_Y(\boldsymbol{I}; \boldsymbol{\theta}_Y) \end{aligned} \tag{1}$$

where a_x and a_y are sampled attribute instances, ϵ is a noise variable which is independent of both X and Y and $h_X(\cdot; \boldsymbol{\theta}_X)$ and $h_Y(\cdot; \boldsymbol{\theta}_Y)$ are the encoders that extract the attribute-specific representations for X and Y, respectively. Under this model, given the distribution of features $\boldsymbol{x} \sim P(\boldsymbol{z}_x)$ and $\boldsymbol{y} \sim P(\boldsymbol{z}_y)$ for the two attributes, we seek to determine whether the attribute-specific representations also follow the same underlying causal relations, i.e., *is* $\boldsymbol{z}_x \to \boldsymbol{z}_y$*?*. The association between these learned attribute features can be well approximated as a post nonlinear causal model (PNL) [63],

$$\boldsymbol{z}_y = f_2(f_1(\boldsymbol{z}_x) + \boldsymbol{\epsilon}) \tag{2}$$

where f_2 and f_1 are non-linear functions with f_2 being continuous and invertible, and $\boldsymbol{\epsilon}$ is a noise variable such that $\boldsymbol{\epsilon} \perp\!\!\!\perp \boldsymbol{z}_x$. The identifiability of the PNL model from observational data was established by Zhang and Hyvärinen [64]. Conceptually, the key idea is that the distribution $P(\boldsymbol{z}_y|\boldsymbol{z}_x)$ in the causal direction is *"less complex"* than that in the anti-causal direction. NCINet, the proposed causal inference approach, is designed to exploit such disparity.

We note that, in the absence of strong assumptions, direct causal relation is indistinguishable from that induced by latent confounding. However, we take inspiration from the ability of humans to accurately infer causal relations only from observations in many cases, and seek to unveil causal motifs between the representations directly from samples.

4. Observational Causal Discovery Problem

Learning-based observational causal discovery considers a dataset $\mathcal{S}$ of n *observational samples*,

$$\mathcal{S} = \{S_i\}_{i=1}^n = \{(\boldsymbol{x}_j, \boldsymbol{y}_j)_{j=1}^{m_i}\}_{i=1}^n \sim P(\boldsymbol{x}, \boldsymbol{y}) \tag{3}$$

where each sample S_i is itself a dataset of m_i representation pairs $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_{m_i}, \boldsymbol{y}_{m_i})\}$, $\boldsymbol{x} \in \mathbb{R}^{d_x}$ and $\boldsymbol{y} \in \mathbb{R}^{d_y}$ are the learned representations corresponding to predicting X and Y respectively, and $P(\boldsymbol{x}, \boldsymbol{y})$ is the joint distribution of the two representations. The joint distribution $P(X, Y)$ can represent different causal relations as shown in Fig. 2, namely, (i) causal class ($X \to Y$); (ii) anti-causal class ($X \leftarrow Y$); (iii) X and Y are unassociated, both in the absence and presence of an unobserved confounder Z.

The key idea of learning-based causal discovery is to exploit the many manifestations of *causal footprints* often present in real-world observational data [48]. For example, oftentimes the functional relationships in the causal direction are "simpler" than those in the anti-causal direction.

[1] The SCEs and the corresponding causal inference queries for the other pairwise causal relations in Fig. 2 can be defined similarly.

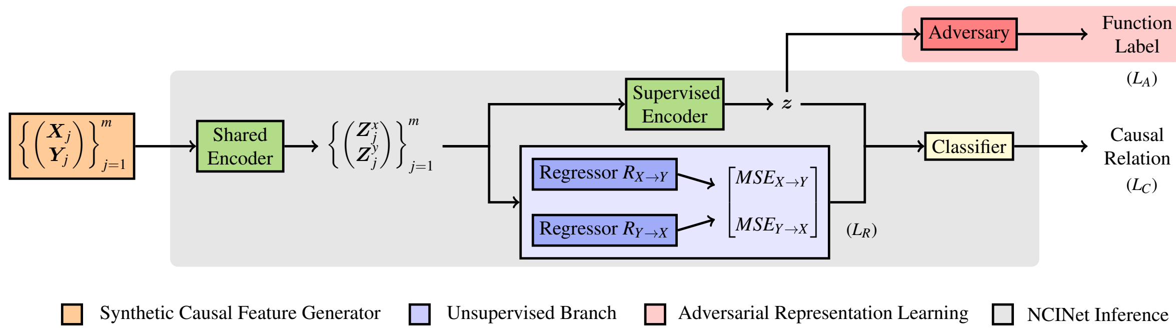


Figure 3. **Overview:** Schematic illustration of Neural Causal Inference Net (NCINet). It comprises of, (1) a shared encoder that maps representations to a common space, (2) a supervised encoder that extracts a representation $\boldsymbol{z}$ from the common space, (3) a causal regression branch that compares the regression errors in the causal and anti-causal direction, (4) an adversary that seeks to extract the function label, and (5) a fusion module that combines information from the two branches and predicts the causal relationship. See text for more details.

Unsupervised methods exploit such *causal signals* either by measuring the complexity of the causal and anti-causal functions [5], the entropy of causal and anti-causal factorizations of the joint distribution [29] or comparing regression errors in the causal and anti-causal direction [4].

Going beyond a specific type of *causal footprint*, supervised methods seek to exploit any and all possible *causal signals* in the observational data by learning to directly predict causal labels from the observation dataset $\mathcal{S}$. *Neural causal models* such as NCC [36], GNN [17] and CE-VAE [38] are a special class of supervised approaches that leverage neural networks based classifiers.

Although both supervised and unsupervised approaches are based on the same principle – namely, *exploiting causal footprints* – they differ in one key aspect. Unlike the unsupervised methods, the supervised approaches need ground truth causal labels to train the causal classifier. However, in most real-world scenarios the ground truth causal graph is unknown. Therefore, the supervised methods are typically trained purely on synthetically generated data and hence suffer from a synthetic-to-real domain generalization gap. Unsupervised methods on the other hand can be applied directly to the observational data of interest and hence are agnostic to the data domain. However, unlike the supervised approaches, the unsupervised methods like RECI [4] exploit only one type of *causal footprint* at a time, e.g., regression errors between the causal and anti-causal directions.

5. Neural Causal Inference Network

Neural Causal Inference Network (NCINet) is a neural causal model for observational causal discovery. Given a pair of high-dimensional attribute-specific representations $S = \{\boldsymbol{x}_j, \boldsymbol{y}_j\}_{j=1}^m$, we seek to determine one of three causal relations, $X \rightarrow Y$, $X \leftarrow Y$ or X is unassociated with Y. Fig. 3 shows a pictorial overview of NCINet along with a causal data generation process that is customized for high-dimensional signals.

Our entire solution is motivated from three perspectives: (1) **Modeling:** As described in the previous section, the supervised and unsupervised models have complementary advantages and limitations. Therefore we incorporate both of them into NCINet to make a final prediction. (2) **Data:** Semantic image attributes (e.g., facial features) span the whole spectrum of pairwise causal relations illustrated in Fig. 2. However, existing supervised and unsupervised learning-based causal discovery methods only consider a subset of these relations (ignoring either the independence class or the unobserved confounder) and are designed for low-dimensional signals, and therefore sub-optimal or insufficient for our purpose. Therefore, we design a synthetic data generation process for obtaining high-dimensional features spanning all possible pairwise causal relations. (3) **Generalization:** For NCINet to generalize from the synthetic training data to real representations we adopt two strategies. First, the synthetic feature generation process includes an ensemble of linear and non-linear causal functions. Second, we employ an adversarial loss to debias the prediction w.r.t the choice of functional classes in the synthetic training data.

NCINet comprises of five components: shared encoder, supervised encoder, causal regression branch, adversary, and classifier. These components are described below.

Encoders: There are two encoders, a shared encoder that maps the pair of representations into an intermediate representation $(\boldsymbol{z}_x, \boldsymbol{z}_y) = (E_{SE}(\boldsymbol{x}), E_{SE}(\boldsymbol{y}))$, and a supervised encoder that extracts features for the final classifier. The latter encoder acts on the concatenated features $\begin{bmatrix}\boldsymbol{z}_x & \boldsymbol{z}_y\end{bmatrix}^T$ and extracts a representation that is average pooled over the m samples in the representation. The resulting feature is denoted as $\boldsymbol{z}$ in Fig. 3.

Causal Regression: The Causal Regression branch of NCINet is inspired by the asymmetry idea proposed by [4], wherein the mean squared error (MSE) of prediction is smaller in the causal direction in comparison to the anti-causal direction, i.e.,

$$\mathbb{E}[(E-\phi(C))^2] \leq \mathbb{E}[(C-\psi(E))^2], \tag{4}$$

where C is the cause and E is the effect, ϕ is the regressor that minimizes the MSE when predicting E from C, and ψ is the regressor that minimizes the MSE when predicting C from E. Therefore, the causal relation can be estimated by comparing the two regression errors. An attractive property of this idea is its inherent ability to generalize to unseen causal data generating functions classes by virtue of being unsupervised and not requiring any learning.

The causal regression branch of NCINet adopt ridge regressors that operates on the intermediate embeddings $(\boldsymbol{z}_x, \boldsymbol{z}_y)$. The causal regressor $R_{X\rightarrow Y}: \boldsymbol{z}_x \mapsto \boldsymbol{y}$ minimizes the MSE, $\frac{1}{m}\sum_{j=1}^{m}\|\hat{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_2^2$, between the prediction $\hat{\boldsymbol{y}} = \left(\boldsymbol{z}_x^T\boldsymbol{z}_x + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{z}_x\boldsymbol{y}$ and ground truth input $\boldsymbol{y}$. Similarly, the anti-causal regressor $R_{X\leftarrow Y}: \boldsymbol{z}_y \mapsto \boldsymbol{x}$ minimizes the MSE, $\frac{1}{m}\sum_{j=1}^{m}\|\hat{\boldsymbol{x}}_j - \boldsymbol{x}_j\|_2^2$, between the prediction $\hat{\boldsymbol{x}} = \left(\boldsymbol{z}_y^T\boldsymbol{z}_y + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{z}_y\boldsymbol{x}$ and ground truth input $\boldsymbol{x}$. The two regressors $R_{X\rightarrow Y}$ and $R_{X\leftarrow Y}$ are trained end-to-end (i.e., we backpropagate through the closed-form ridge regressor solution) along with the rest of the components in NCINet. Therefore, the loss from the regression branch is,

$$L_R = \frac{1}{m}\sum_{i=1}^{m}\|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\|_2^2 \tag{5}$$

Adversarial Loss: The features $\boldsymbol{z}$ extracted from the supervised encoder potentially still contain information specific to the function class that generated the synthetic features. However, the generalization performance of NCINet may be hampered if the downstream classifier exploits any spurious correlation between the function class-specific information and the ground truth causal relations. Therefore, we measure the amount of information in $\boldsymbol{z}$ about the function class through an adversary and minimize it. This type of adversary is typically modeled as a neural network and optimized via min-max optimization, which can be unstable in practice [13,22]. For ease of optimization we instead model the adversary by a kernel ridge-regressor which admits a closed-form solution $\hat{\boldsymbol{y}}_f = \boldsymbol{K}\left(\boldsymbol{K}+\beta\boldsymbol{I}\right)^{-1}\boldsymbol{y}_f$, where $\boldsymbol{y}_f$ is the one-hot vector representing the function class of the synthetic data, β is a regularization, and $\boldsymbol{K}$ is a kernel matrix computed from the features $\boldsymbol{z}$. The loss from the adversary which we backpropagate through is,

$$L_A = -\|\boldsymbol{y}_f - \hat{\boldsymbol{y}}_f\|_2^2 = -\|\boldsymbol{y}_f - \boldsymbol{K}\left(\boldsymbol{K}+\beta\boldsymbol{I}\right)^{-1}\boldsymbol{y}_f\|_2^2 \tag{6}$$

Classifier: Finally, the supervised classifier concatenates the features $\boldsymbol{z}$ from the supervised encoder with the output $\left[MSE_{X\rightarrow Y} \quad MSE_{Y\rightarrow X} \quad \frac{min(MSE_{X\rightarrow Y}, MSE_{Y\rightarrow X})}{max(MSE_{X\rightarrow Y}, MSE_{Y\rightarrow X})})\right]^T$ of the causal regressors, and categorizes the causal relations into three classes as follows.

$$l = \begin{cases} 0 & \text{if } X \text{ unassociated with } Y \text{ i.e., no causal relation} \\ 1 & \text{if } X \rightarrow Y \\ 2 & \text{if } X \leftarrow Y \end{cases}$$

We note that this is unlike existing methods for causal inference such as NCC [36], RECI [4], etc., which only categorize the causal relations into two categories, namely causal and anti-causal. However, in many practical scenarios, the image attributes and their corresponding representations could be very weak as we discuss in Section 7.2. The classifier minimizes the cross-entropy loss L_C between its prediction and the ground truth causal relations l.

All the components of NCINet are trained end-to-end by simultaneously optimizing all the intermediate losses i.e., $Loss = L_C + L_R + \lambda L_A$, where λ is the weight associated with the adversarial loss. The classifier will learn to exploit all the *causal footprints* in the data aided by the causal regressors and the adversary. The features from the causal regressors help exploit the *causal footprint* corresponding to the difference in regression errors, while the adversary helps the classifier to reduce the synthetic-to-real domain generalization gap. Interaction between the regressors, adversary, and the final classifier is induced by the common intermediate representation space $(\boldsymbol{z}_x, \boldsymbol{z}_y)$, on which all of them operate.

6. Experiments: Neural Causal Inference

In this section, we evaluate the performance and generalization ability of NCINet in comparison to existing baseline methods on synthetically generated high-dimensional representations with known causal relations.

Data and Training: The lack of large-scale datasets with ground truth causal labels precludes causal discovery models from being trained on real-world observational data. Therefore, it is standard practice to train and evaluate causal discovery models on synthetic observational data. Models trained in this manner can now be applied directly to real-world observational data. Synthetic data generation typically follows the additive noise model [46], where an effect variable is obtained as a function of causal variable and perturbed with independent additive noise. We adopt the same additive noise model as our causal mechanism.

To improve generalization capability, we diversify the synthetic training data. Specifically, we adopt an ensemble of different high-dimensional causal functions including, Linear, Hadamard, Bilinear, Cubic Spline, and Neural Networks. See the supplementary material for more details. In each training epoch, we generate 1000 samples, where each data sample consists of 100 feature pairs (i.e., $m = 100$) by randomly sampling one of the causal functions and their respective parameters. Integrating data generation into the training process ensures that the models learn from an infinite stream of non-repeating data.

We generate the pairs of representations $(\boldsymbol{x}, \boldsymbol{y})$ via ancestral sampling. For example, in the case of $\mathcal{G}_1$ where $\boldsymbol{x} \rightarrow \boldsymbol{y}$, the synthetic representations are generated as follows, $P(\boldsymbol{w}) \rightarrow P(\boldsymbol{x}|\boldsymbol{w}) \rightarrow P(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{w})$, where $\boldsymbol{w}$ accounts for all the unobserved confounders. More details can be found in the supplementary material.

Baselines: We consider four baseline methods, ANM [20], Bivariate Fit (BFit) [23], NCC [36] and RECI [4]. These methods, however, were originally designed for causal inference on one-dimensional variables and to distinguish between causal and anti-causal directions. Therefore, we extend them to high-dimensional data, as well as to distinguish between causal direction, anti-causal direction, and no causal relation. Specifically, for NCC, we concatenate high-dimensional features $\boldsymbol{x}$ and $\boldsymbol{y}$ as the input to the network and change the output layer to three classes. For the unsupervised methods, ANM, BFit, and RECI, we regress directly on high-dimensional features $\boldsymbol{x}$ and $\boldsymbol{y}$ as required. Since these methods are score based i.e. $score > 0$ represents the causal direction and $score < 0$ represents the anti-causal direction we introduce an additional threshold to identify the no causal relation case i.e. if $|score| < threshold$. We use a separate validation set to determine the optimal threshold for each unsupervised method.

Generalization Results: To evaluate the performance of the models and their generalization ability, we adopt a leave-one-function out evaluation protocol. For evaluating each causal function, we train the models using data generated by all the other causal functions across all the causal graphs in Fig. 2. The results are shown in Table 1 for 8 dimensional representations. We observe that overall NCINet outperforms all the baselines.

Table 1. Leave-one-function out mean accuracy (%) of five runs on different causal functions with 8 dimensional features (see supplementary for more details). Best results are in bold.

Methods	Linear	Hadamard	Bilinear	Cubic Spline	NN	Average
ANM [20]	31.87	32.49	32.94	33.66	33.08	32.81
Bfit [23]	34.89	54.76	53.69	**77.79**	38.26	51.88
NCC [36]	52.64	83.93	85.66	77.03	56.56	71.16
RECI [4]	42.73	**89.66**	**92.02**	71.49	60.23	71.43
NCINet	**64.16**	81.13	89.73	71.33	**69.53**	**75.17**

7. Causal Inference on Learned Features

Our goal in this section is to estimate the causal relation between attribute-specific learned representations and verify if it is consistent with the causal relations between the corresponding labels. Furthermore, we would like to perform this analysis for all the different types of causal relations shown in Fig. 2. However, real-world datasets only provide a fixed collection of images and their corresponding labels, and as such do not afford any explicit control over the type and strength of causal relations between the attributes. To overcome the aforementioned limitations we consider causal inference on learned representations under two scenarios where the causal relations between the attributes are known and unknown.

Datasets: We consider three image datasets with multiple multi-label attributes; (1) 3D Shapes [7] which contains 480,000 images. Each image is generated from six latent factors (floor hue, wall hue, object hue, scale, shape, and orientation) which serve as our image attributes. (2) CASIA WebFace [60] which contains 494,414 images of 10,575 classes. While the dataset does not come with attribute annotations, we annotate each image with eight multi-label attributes[2] (color of hair, visibility of eyes, type of eyewear, facial hair, whether the mouth is open, smiling or not, wearing a hat, visibility of forehead, and gender). The annotations for this dataset have been made publicly available to the research community. (3) CelebA [34] which contains 202,599 images, each of which is annotated with 40 binary attributes. However, this dataset suffers from severe class imbalance across most attributes. Experimental results on this dataset can be found in the supplementary material.

Learning Attribute-Specific Representations: We learn attribute-specific representations by learning attribute predictors for each attribute. For predicting the attributes, we use a 5-layer CNN and ResNet-18 [18] for the 3D Shapes and CASIA-WebFace datasets, respectively. The attribute predictors are optimized using AdamW [37] with a learning rate of 5×10^{-4} and a weight decay of 5×10^{-4}. Upon convergence of the attribute predictor training, we use the trained model to extract representations from the layer before the linear classifier at the end.

Causal Inference Baselines: To infer the causal relations between the learned attribute representations we apply NCINet and two other baselines NCC and RECI. Both NCINet and NCC are trained on the synthetic dataset as described in Section 6.

7.1. Known Causal Relations Between Labels

In this experiment, we resample the datasets to obtain samples with the desired type of causal relations. Consequently, in this scenario, the causal relationship between the attribute labels is known, and we seek to verify if the corresponding attribute-specific learned representations also satisfy the same causal relations. We note that, although the data generated in this way will not reflect the true underlying causal relations between the attributes, it nonetheless al-

[2]The choice of attributes and labels for each may arguably still not fully reflect the real-world. Nonetheless, we believe this dataset could be a valuable resource for causal analysis tasks.

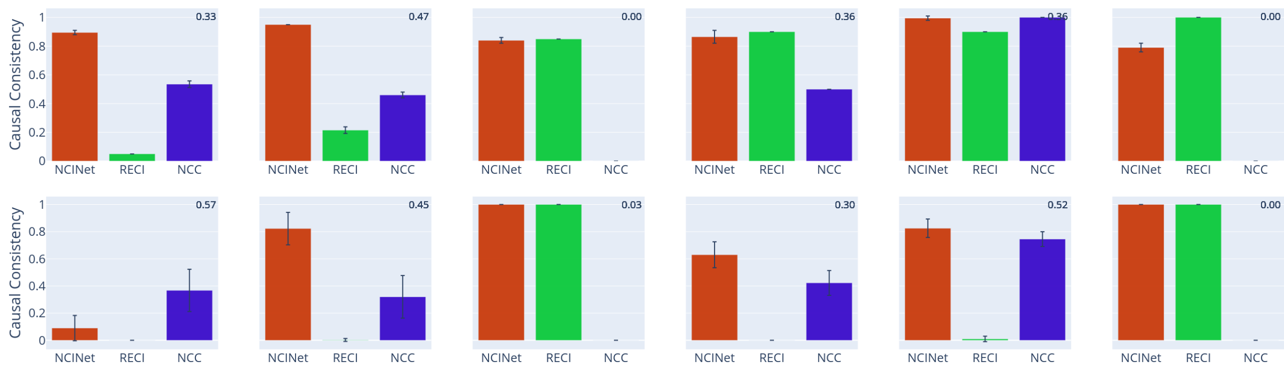


Figure 4. *Causal consistency* between the labels and representations of attributes. The columns represent the different types of causal relations starting from $\mathcal{G}_1$ on the left and $\mathcal{G}_6$ on the right. The strength of the causal relation between the labels, estimated through [29], is shown in top right-hand corner of each subplot. (Top) 3D Shape and (Bottom) CASIA WebFace dataset.

lows us to perform controlled experiments. We chose floor hue and wall hue as the attribute for 3D Shapes, and visibility of the forehead, and whether the mouth is open as the attributes for CASIA-WebFace. The choice of the attributes of CASIA-WebFace was motivated by the fact that these two attributes were the most sample balanced pair of attributes. For each type of causal relation in Fig.2, we sample 2000/2000 and 8000/2000 images for training/testing on 3D Shapes and CASIA-WebFace, respectively.

Generating Images with Causally Associated Attributes: The data sampling process proceeds in two phases. In the first phase, we generate the attribute labels with known causal relations. We represent each type of causal graph via its corresponding Bayesian Network with hand-designed conditional probability tables. Then we sample batches of labels with known causal relations through Gibbs Sampling. To further ensure that the sampled labels correspond to the desired causal relation, we measure the strength of their causal relation through an entropic causal inference method [29]. In the second phase, we sample images that conform to the sampled attribute labels. On the 3D Shapes dataset, to increase the diversity of the images and ensure that the representation learning task is sufficiently challenging, the images are corrupted with one of three types of noise, Gaussian, Shot, or Impulse.

Results: To measure the consistency between the causal relations of the labels and the causal relations of the respective representations, we introduce a new metric dubbed *causal consistency*. For a given set of learned representations $(\boldsymbol{x}_j, \boldsymbol{y}_j)_{j=1}^{m}$, we split it into multiple non-overlapping subsets. The causal relation is estimated for each subset and we measure how many of them are consistent with the causal relation l between the labels. Evaluating over multiple subsets serves to prevent outliers from severely affecting the causal inference estimates (see supplementary material for more details). Fig. 4 shows the *causal consistency* results across the different types of causal relations. We make the following observations: (1) In most cases, across both 3D Shapes and CASIA-WebFace, the causal relationship between the learned representations is highly consistent with that of the labels. This empirical evidence is encouraging since it suggests that representation learning algorithms are capable of mimicking the causal relations inherent to the training data. (2) In the controlled setting of this experiment, among the three causal inference methods, NCINet appears to provide more stable and consistent estimates of causal relations across the different causal graphs and datasets, followed by RECI and NCC.

7.2. Unknown Causal Relations Between Labels

In this experiment, we consider the original CASIA-WebFace dataset as is, without any controlled sampling. We choose smiling or not and visibility of eyes as the two attributes to investigate. The attribute predictors are trained/validated on 10,000/10,000 randomly sampled images using a ResNet-18 architecture. Other training details are similar to the experiment in Section 7.1. Since the true causal relation between the labels is unknown, we use an entropic causal inference method [29] to estimate it. While the causal relation between the labels suggests that smiling has an effect on the visibility of eyes its causal strength is very weak (0.23/0.20 for training/validation). The causal relation between the learned representations follow the same trend, with around 20% of the representation subsets agreeing with smiling having an effect on the visibility of eyes, while 80% of them suggest that there is no causal relation between the two attributes.

8. Discussion

This section analyzes the effect of various aspects of representation learning on the *causal consistency* between the learned representations and the labels.

Effect of Adversarial Debiasing on NCINet: We study the contribution of adversarial loss L_A on the generalization

Table 2. Effect of Adversarial Debiasing on NCINet (one run)

NCINet	Linear	Hadamard	Bilinear	Cubic spline	NN	Average
w/o Adv	66.50	80.33	89.67	70.5	67.17	74.83
w/ Adv	66.67	80.50	90.17	71.00	68.33	75.33

performance of NCINet. Table 2 reports the causal inference results on the synthetically generated representations in Section 6. Overall the adversarial loss aids in improving the generalization ability of NCINet, thereby validating our hypothesis that the features z from the supervised encoder still contains information specific to the function class.

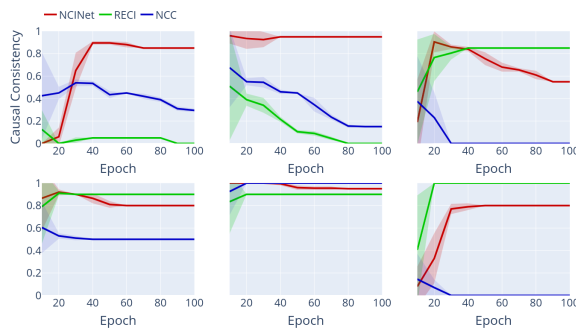


Figure 5. *Causal consistency* as a function of training epochs for the six different types of causal relations. (Top) $\mathcal{G}_1$-$\mathcal{G}_3$ and (Bottom) $\mathcal{G}_4$-$\mathcal{G}_6$. Bands show standard deviation.

Effect of Training Epochs: Here we study the causal relations between the representations as a function of the training epochs. In this experiment, we test the causal consistencies of features extracted by models from each training epoch, and show the average value of every 10 epochs. We hypothesize that as the attribute prediction performance of the representations improves, the causal relation between the representations will also become more consistent with the causal relation between the labels. Fig. 5 shows the *causal consistency* as a function of the training epochs. We make three observations: (1) as training progresses, the *causal consistency* of NCINet improves and remains stable which is consistent with our hypothesis. (2) Although RECI fails on $\mathcal{G}_1$ and $\mathcal{G}_2$, it follows the same trend in the other cases. (3) NCC is prone to failure in no causal relation data.

Effect of Overfitting: Here we study the effect of overfitting on the *causal consistency* between the representations. We hypothesize that as the representation learning process overfits, the *causal consistency* on the validation features will drop. Fig. 6 shows the results of this experiment. We observe that after overfitting, the *causal consistency* drops for features from both the training and test set. However, the former still retain some *causal consistency*.

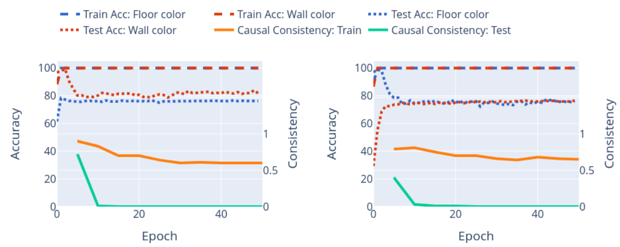


Figure 6. Effect of overfitting on causal consistency of NCINet.

9. Conclusion

This paper sought to answer the following questions: *Do learned attribute-specific representations also satisfy the same underlying causal relations? And, if so, to what extent?* To answer these questions, we designed Neural Causal Inference Network (NCINet) for causal discovery from high-dimensional representations. By bringing together ideas from learning-based supervised, unsupervised causal prediction methods and adversarial debiasing, NCINet exhibits significantly better causal inference generalization performance. We applied NCINet to estimate the *causal consistency* between learned representations and the underlying labels in two scenarios, one where the causal relations are known through controlled sampling, and the other where the causal relationships are unknown. Furthermore, we analyzed the effect of overfitting and training epochs on the *causal consistency*. Our experimental results suggest that learned attribute-specific representations *indeed* satisfy the same causal relations between the corresponding attribute labels under controlled scenarios and with high causal strength.

Causal analysis of learned representations is a novel, challenging, and important task. Our work presents a solid yet preliminary effort at answering the questions raised in this paper. Our work is limited from the perspective that there exist many potentially interesting and related aspects of this problem that we did not explore here. From a technical perspective, we foresee two limitations of our work. (1) Unlike the unsupervised methods like RECI, NCINet needs to be retrained if the dimensionality of the representation changes. (2) NCINet and the baselines exhibit poor causal inference performance on data with weak causal relations. As the causal relation gets weaker, it is increasingly difficult to distinguish it from the no causal association case. Furthermore, our data generating process does not afford explicit control over the causal strength.

Acknowledgements: This work was performed under the following financial assistance award 60NANB18D210 from U.S. Department of Commerce, National Institute of Standards and Technology.

References

[1] Timo Ahonen, Abdenour Hadid, and Matti Pietikäinen. Face recognition with local binary patterns. In *European Conference on Computer Vision (ECCV)*, 2004. 2

[2] Bryon Aragam and Qing Zhou. Concave penalized estimation of sparse gaussian bayesian networks. *The Journal of Machine Learning Research*, 16(1):2273–2328, 2015. 2

[3] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019. 2

[4] Patrick Blöbaum, Dominik Janzing, Takashi Washio, Shohei Shimizu, and Bernhard Schölkopf. Cause-effect inference by comparing regression errors. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018. 1, 2, 4, 5, 6

[5] Kailash Budhathoki and Jilles Vreeken. Origo: causal inference by compression. *Knowledge and Information Systems*, 56(2):285–307, 2018. 2, 3, 4

[6] Peter Bühlmann, Jonas Peters, Jan Ernest, et al. Cam: Causal additive models, high-dimensional order search and penalized regression. *The Annals of Statistics*, 42(6):2526–2556, 2014. 2

[7] Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018. 2, 6

[8] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 2

[10] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 2

[11] David Maxwell Chickering. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554, 2002. 2

[12] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005. 2

[13] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 5

[14] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[15] José AR Fonollosa. Conditional distribution variability measures for causality detection. In *Cause Effect Pairs in Machine Learning*, pages 339–347. Springer, 2019. 3

[16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[17] Olivier Goudet, Diviyan Kalainathan, Philippe Caillou, Isabelle Guyon, David Lopez-Paz, and Michele Sebag. Learning functional causal models with generative neural networks. In *Explainable and Interpretable Models in Computer Vision and Machine Learning*, pages 39–80. Springer, 2018. 2, 4

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 6

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017. 2

[20] Patrik O Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems (NeurIPS)*, 2009. 2, 6

[21] Dominik Janzing, Patrik O Hoyer, and Bernhard Schölkopf. Telling cause from effect based on high-dimensional observations. In *International Conference on Machine Learning (ICML)*, 2010. 3

[22] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? *arXiv preprint arXiv:1902.00618*, 2019. 5

[23] Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*, 2019. 6

[24] Diviyan Kalainathan, Olivier Goudet, Isabelle Guyon, David Lopez-Paz, and Michèle Sebag. Structural agnostic modeling: Adversarial learning of causal graphs. *arXiv preprint arXiv:1803.04929*, 2018. 2

[25] Diviyan Kalainathan, Olivier Goudet, Michèle Sebag, and Isabelle Guyon. Discriminant learning machines. *Cause Effect Pairs in Machine Learning*, pages 155–189, 2019. 3

[26] Markus Kalisch and Peter Bühlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8(Mar):613–636, 2007. 2

[27] Nan Rosemary Ke, Olexa Bilaniuk, Anirudh Goyal, Stefan Bauer, Hugo Larochelle, Bernhard Schölkopf, Michael C Mozer, Chris Pal, and Yoshua Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019. 2

[28] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning (ICML)*, 2018. 2

[29] Murat Kocaoglu, Alexandros G Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 2, 4, 7

[30] Murat Kocaoglu, Sanjay Shakkottai, Alexandros G Dimakis, Constantine Caramanis, and Sriram Vishwanath. Entropic

latent variable discovery. *arXiv preprint arXiv:1807.10399*, 2018. 2

[31] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. Causalgan: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[33] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song. Sphereface: Deep hypersphere embedding for face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015. 2, 6

[35] David Lopez-Paz, Krikamol Muandet, Bernhard Schölkopf, and Iliya Tolstikhin. Towards a learning theory of cause-effect inference. In *International Conference on Machine Learning (ICML)*, 2015. 3

[36] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 3, 4, 5, 6

[37] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018. 6

[38] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 2, 4

[39] David G Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision (ICCV)*, 1999. 2

[40] Alexander Marx and Jilles Vreeken. Causal inference on multivariate and mixed-type data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, 2018. 3

[41] Joris M Mooij, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. Distinguishing cause from effect using observational data: methods and benchmarks. *The Journal of Machine Learning Research*, 17(1):1103–1204, 2016. 2

[42] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision (ECCV)*, 2016. 2

[43] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[44] Judea Pearl. *Causality: models, reasoning and inference*, volume 29. Springer, 2000. 2

[45] Judea Pearl. *Causality*. Cambridge university press, 2009. 2

[46] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 2000. 5

[47] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332*, 2015. 3

[48] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *The Journal of Machine Learning Research*, 15(1):2009–2053, 2014. 3

[49] Francesco Pittaluga, Sanjeev J Koppal, Sing Bing Kang, and Sudipta N Sinha. Revealing scenes by inverting structure from motion reconstructions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[50] Proteek Chandan Roy and Vishnu Naresh Boddeti. Mitigating information leakage in image representations: A maximum entropy approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[51] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000. 2

[52] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 4, page 12, 2017. 2

[53] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 3

[54] Luan Tran, Xi Yin, and Xiaoming Liu. Disentangled representation learning gan for pose-invariant face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[55] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1991. 2

[56] Jilles Vreeken. Causal inference by direction of information. In *SIAM International Conference on Data Mining ((SDM))*, 2015. 2, 3

[57] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 3

[58] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021. 3

[59] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9847–9857, 2021. 3

[60] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 2, 6

[61] Zhongqi Yue, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Transporting causal mechanisms for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8599–8608, 2021. 3

[62] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3

[63] Kun Zhang and Aapo Hyvärinen. Distinguishing causes from effects using nonlinear acyclic causal models. In *Causality: Objectives and Assessment*. PMLR, 2010. 3

[64] Kun Zhang and Aapo Hyvarinen. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012. 3