Do learned representations respect causal relationships?
(Supplementary Material)

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

In this supplementary material, we include,

1. Precise description and definition of causal consistency in Section 1
2. Experimental results of causal consistency on CelebA Face Dataset in Section 2.
3. An ablation experiment on effect of the adversarial loss on the performance of NCINet in Section 3.
4. Additional experimental results analyzing the effect of factors such as representation dimensionality and network architecture for learning the representations on *causal consistency* in Section 4.
5. Details of the process for generating the synthetic representation for training NCINet and the baselines in Section 5.
6. Details of the process for generating images with causally associated attributes in Section 6.
7. Details of facial attribute annotation on the CASIA dataset used for our experiments in Section 7.

1. Definition of causal consistency

Datasets are divided into subsets. Causal consistency is the ratio of subsets whose causal relation between representations matches that of the labels, with higher values representing higher consistency. Further, we compute average causal consistency (and confidence intervals) across a small interval K (ten) of epochs after representation learning has converged. Overall, Causal consistency $= \frac{1}{K}\sum_{k=1}^{K}\frac{\#\text{consistent subsets}}{\#\text{subsets}}$.

2. Causal consistency of CelebA

We also conduct causal inference on representations learned on the CelebA dataset. Specifically, we experiment on the case where causal relations between labels are *unknown*. Similar to the experiments on the CASIA dataset, we chose smiling and narrow eyes as the two attributes to investigate, train and validate the attribute predictors on 10,000/10,000 randomly sampled images using a ResNet-18 architecture. We also apply the entropic causal inference method [5] to estimate the causal relation between labels and finding that smiling is a cause of narrow eyes. Table 1 shows the causal inference results of NCINet and two baseline. NCINet exhibits strong *causal consistency* in the correct causal direction. Due to the challenge of selecting a score threshold (see Section 6 of main paper for details) for RECI that generalizes beyond the training data, it classifies all sample as no causal relation. However, if we set the threshold to 0 and let RECI only infer causal and anti-causal direction, the majority samples will also be inferred as the same directions with labels, which shows that in this case, the causal relation between the features is indeed consistent with that of the labels.

Table 1. Causal consistency on CelebA.

	NCINet	RECI	NCC
Causal Consistency	0.82	0.00	0.01

3. Ablation: Effect of ARL

To investigate how adversarial loss contributes to NCINet, we test three different λ values in $Loss = L_C + L_R + \lambda L_A$ and present their generalization results on high-dimensional synthetic data . Table 2 shows the generalization results of using different adversarial weight. The results indicate that for data generated from different causal functions, the optimal weight λ is different. However, even a small weight of ARL loss could help the model's generalization ability.

Figure 1 shows different components of training loss. With a wight λ associated with the adversarial loss, all losses are roughly of the same order of magnitude and well balanced.

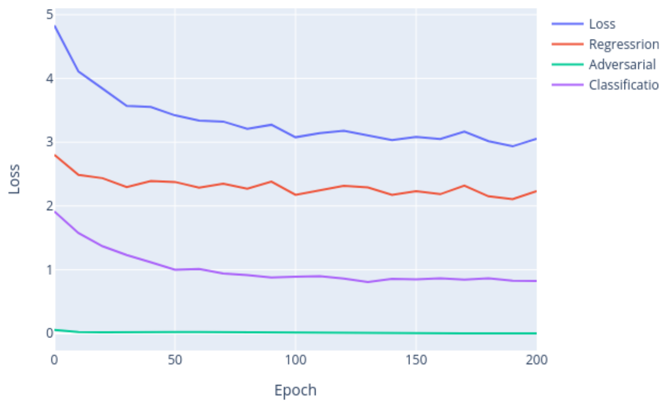

Figure 1. Different components of training loss

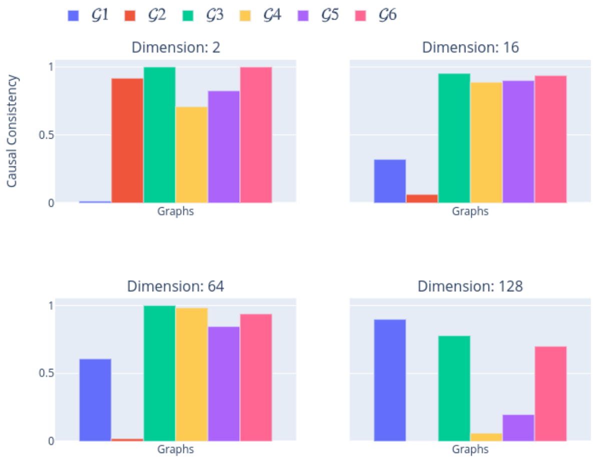

Figure 2. Causal consistency and feature dimension

Table 2. Effect of Adversarial Debiasing on Weight (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
optimal Adv	66.67	80.50	90.17	71.00	68.33	75.33
λ=0.5	66.67	79.67	89.83	70.83	68.33	75.06
λ=2	66.67	79.83	89.67	70.83	68.33	75.06
λ=10	65.00	80.50	90.17	71.00	68.17	74.96

Table 3. Sample complexity ablation. We used $m = 100$ for experiments in paper. (one run)

	Linear	Hadamard	Bilinear	Cubic Spline	NN	Average
m=10	56.50	37.00	30.67	34.00	33.67	38.36
m=100	66.67	80.50	90.17	71.00	68.33	75.33
m=1000	58.83	81.83	90.17	70.33	66.33	73.49

4. Discussion

Effect of Representation Dimensionality: To investigate the effect of representation dimensionality on the inherent causal relations, we evaluate *causal consistency* across different representation dimensionalities on the CASIA WebFace dataset. We set different number of dimensions for the layer before the last linear classifier in the attribute predictor, and extract representations from models that are trained to convergence. Figure 2 shows the *causal consistency*. We observe that there is slight degradation in the *causal consistency* as the number of dimensions increases, especially at 128 dimensions. However, a more careful and controlled experiment is necessary in order to gain a deeper understanding on the role of representation dimensionality on *causal consistency*.

Effect of Architecture : Here we seek to understand if the network architecture has an effect on the causal relations between learned attributes. Therefore, we use four different architecture, including ResNet18, ResNet34, ResNet50 and WideResNet as the attribute predictor for Casia Dataset. Figure 3 shows the causal consistency for multiple network architectures. The results indicates that changes in network architecture have a larger impact on $\mathcal{G}_1$ and $\mathcal{G}_2$, while providing more stable results on other graphs.

Effect of Sample Complexity : We also study the effect of sample complexity. We set different sample size m and verify the generalization performance. As shown in Table 3, as sample size increases the results generalize better but plateau with a certain size of sample complexity. The results indicates that to infer the causal relation an adequate number of pairs are needed for each sample.

Results of Multiple Runs: To evaluate the stability and effectiveness of different methods, we run all baselines for five times in the leave-one-function-out generalization experiment, and present their mean accuracy and standard deviation. Specifically, in each run, we generate five different testing datasets for each causal function. The results, shown in Table 4, indicate that NCINet have a more stable result comparing with other baselines.

Standard Deviation: Table 5a and 5b show mean and standard deviation (specific numbers of Figure 4 in main paper) over the small interval of epochs after representation learning has converged on the 3d shape and Casia datasets. As can be observed that causal consistency of NCINet, from one epoch to the other is very stable, which is comparable to unsupervised method.

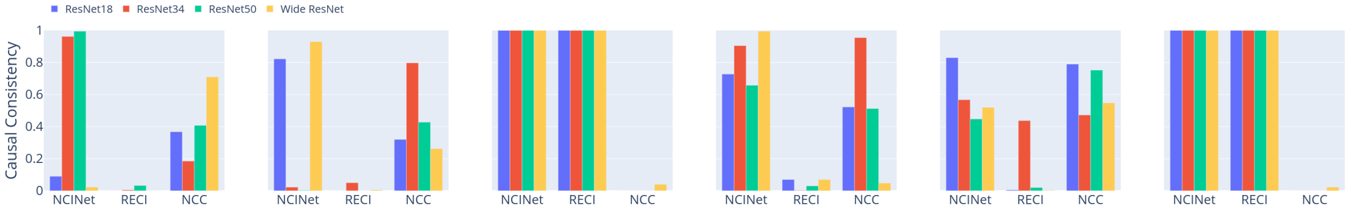

Figure 3. Effect of Architecture and Model Size. From left to right, the plots represent the causal relations encoded by $\mathcal{G}_1$ to $\mathcal{G}_6$.

Table 4. Leave-one-function out accuracy (%) on different causal functions of different runs.

Methods	Linear	Hadamard	Bilinear	Cubic Spline	NN	Average
ANM [3]	31.87 ± 1.55	32.49 ± 2.31	32.94 ± 0.72	33.66 ± 2.69	33.08 ± 1.15	32.81 ± 1.68
Bfit [4]	34.89 ± 2.01	54.76 ± 1.03	53.69 ± 1.70	$\mathbf{77.79 \pm 2.40}$	38.26 ± 1.32	51.88 ± 1.70
NCC [7]	52.64 ± 2.79	83.93 ± 1.55	85.66 ± 1.76	77.03 ± 1.42	56.56 ± 1.37	71.16 ± 1.78
RECI [1]	42.73 ± 1.46	$\mathbf{89.66 \pm 1.50}$	$\mathbf{92.02 \pm 1.01}$	71.49 ± 0.79	60.23 ± 2.15	71.43 ± 1.38
NCINet	$\mathbf{64.16 \pm 2.33}$	81.13 ± 0.70	89.73 ± 0.71	71.33 ± 0.33	$\mathbf{69.53 \pm 0.94}$	$\mathbf{75.17 \pm 1.00}$

(a) Causal consistency on 3Dshape with standard deviation

	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	$\mathcal{G}_6$
NCINet	0.89 ± 0.01	0.94 ± 0.00	0.83 ± 0.02	0.86 ± 0.04	0.99 ± 0.01	0.79 ± 0.03
RECI	0.05 ± 0.00	0.21 ± 0.02	0.85 ± 0.00	0.90 ± 0.00	0.90 ± 0.00	1.00 ± 0.00
NCC	0.53 ± 0.02	0.46 ± 0.02	0.00 ± 0.00	0.50 ± 0.00	1.00 ± 0.00	0.00 ± 0.00

(b) Causal consistency on Casia with standard deviation

	$\mathcal{G}_1$	$\mathcal{G}_2$	$\mathcal{G}_3$	$\mathcal{G}_4$	$\mathcal{G}_5$	$\mathcal{G}_6$
NCINet	0.09 ± 0.09	0.82 ± 0.11	1.00 ± 0.00	0.63 ± 0.09	0.82 ± 0.06	1.00 ± 0.00
RECI	0.00 ± 0.00	0.00 ± 0.01	1.00 ± 0.00	0.00 ± 0.00	0.01 ± 0.02	1.00 ± 0.00
NCC	0.36 ± 0.15	0.32 ± 0.01	0.00 ± 0.00	0.42 ± 0.09	0.74 ± 0.05	0.00 ± 0.00

5. Synthetic Causal Representation Generating Process

The following steps are the detailed data generation process. In this illustration, we taking the case of X being the cause variable for example:

- **Generating initial cause data**: we first sample initial data W from a mixture of Gaussian distributions, and then generate synthetic representation X through a causal function: $X = f(W) + \boldsymbol{\epsilon}$.
- **Generating ground truth label**: Randomly select one of the first six scenarios in Figure 2 of main paper, and assign the corresponding label to l.
- **Generating high-dimensional causal relation**: Randomly select one of the five high-dimensional causal function to establish causal relation from cause to effect: $Y = f(X) + \boldsymbol{\epsilon}$.
- **Confounder Cases**: In the cases which involves confounder Z (e.g., $\mathcal{G}_4$), we first establish the causal relation of $Z \to X$, and then establish the causal relation of $X, Z \to Y$: $Y = f(X, Z) + \boldsymbol{\epsilon}$. In the cases where X and Y have no causal relation (i.e. $l = 0$), if it involves confounder Z, we establish the causal relation of $Z \to X$ and $Z \to Y$,if not, we leave X and Y as their initial values.

The five high-dimensional causal functions are specified in Table 6, with both w/o confounder and w/ confounder cases. For linear and quadratic functions, we directly multiply the cause variable with coefficient matrices in their form. For Bilinear function, we apply a bilinear transformation to the cause variable. For cubic spline function, we follow [7], applying a cubic Hermite spline function. We draw k knots from $\mathcal{N}(0, 1)$, where k is drawn from RandomInteger(5, 20). For Neural Networks function, we apply multilayer perceptrons with hidden layers and numbers of hidden neurons drawn from RandomInteger(0, 3) and RandomInteger(8, 20). For each function, its parameters (e.g., $\boldsymbol{A}$, $\boldsymbol{B}$ or MLP weights) are drawn at random from $\mathcal{N}(0, 1)$ for each data sample. The noise terms $\boldsymbol{\epsilon}$ are sampled from Gaussian(0, v), where $v \sim$ Uniform(0, 0.1). After each operation, including data initialization and causal relation establishment, the data will be normalized to zero mean and unit variance. Note that for initial data generating, we also apply same causal function as high-dimensional causal relation generating.

6. Generating Images with Causally Associated Attributes

As mentioned in Section 7 of the main paper, the image generating process contains two phases. In the first

Table 6. Generative Model for Synthetic Causal Representations

Causal functions	Linear	Hadamard	Bilinear	Cubic spline	NN
w/o Confounder	$\boldsymbol{w} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and $\boldsymbol{x} = g_x(\boldsymbol{w}) + \boldsymbol{\epsilon}$				
	$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$	$\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{x} \odot \boldsymbol{x}) + \boldsymbol{B}\boldsymbol{x} + \boldsymbol{\epsilon}$	$\boldsymbol{y} = \boldsymbol{x}^{\boldsymbol{T}} \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\epsilon}$	$\boldsymbol{y} = Spline(\boldsymbol{x}) + \boldsymbol{\epsilon}$	$\boldsymbol{y} = MLP(\boldsymbol{x}) + \boldsymbol{\epsilon}$
w/ Confounder	$\boldsymbol{w} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ $\quad \boldsymbol{z} = g_z(\boldsymbol{w}) + \boldsymbol{\epsilon}$ $\quad \boldsymbol{x} = g_x(\boldsymbol{z}) + \boldsymbol{\epsilon}$				
	$\boldsymbol{y} = \boldsymbol{A}\tilde{\boldsymbol{z}} + \boldsymbol{\epsilon}$	$\boldsymbol{y} = \boldsymbol{A}(\tilde{\boldsymbol{z}} \odot \tilde{\boldsymbol{z}}) + \boldsymbol{B}\tilde{\boldsymbol{z}} + \boldsymbol{\epsilon}$	$\boldsymbol{y} = \tilde{\boldsymbol{z}}^{T} \boldsymbol{A}\tilde{\boldsymbol{z}} + \boldsymbol{\epsilon}$	$\boldsymbol{y} = Spline(\boldsymbol{x}) + Spline(\boldsymbol{z}) + \boldsymbol{\epsilon}$	$\boldsymbol{y} = MLP(\tilde{\boldsymbol{z}}) + \boldsymbol{\epsilon}$

$\tilde{\boldsymbol{z}}$ indicates concatenation of $\boldsymbol{x}$ and $\boldsymbol{z}$.

phases, we sample labels with six causal relations of Figure 2 in main paper. We first build Bayesian Network with hand-designed conditional probability tables of six causal graphs, and then conduct Gibbs Sampling to get attribute labels with known causal relation. The goal of the second phase is to sample images using the labels with known causal relation. For example, in 3D Shapes Dataset, we select attribute floor hue and wall hue as the attribute X and Y in six causal graphs. Then we sample images according to the labels with known causal relations, that is, we select images whose attribute floor hue and wall hue are same with the sampled labels, while we keep other attributes random. For each image, we also randomly add one of three types of noise, Gaussian, Shot, or Impulse. Figure 4 shows examples of images generated from the 3D Shapes dataset. Similarly for facial dataset CelebA and Casia Dataset, we also apply same strategy to sample images using labels with known causal relationship from original dataset.

7. Facial Attribute Annotations

Progress in causal discovery methods for computer vision has been hampered by the lack of a large-scale dataset annotated with different underlying causal relations. We posit that existing datasets such as CelebA [6], which has annotations of multi-label attributes in the form of binary labels, is inadequate for causal discovery for a couple of reasons. First, a majority of the images for each attribute are highly imbalanced towards one of the two classes. And more importantly, we observed that a majority of the binary labels are very close to being independent of each other. As such, it may not accurately reflect the causal relations in the real-world and are for the most part are unsuitable as an evaluation benchmark.

To overcome this hurdle we adopt the CASIA-Webface [8] dataset, a large public face dataset with 10,575 people and 494,414 images in total, for our experiments. Since this dataset is designed for face verification and recognition problems, only identity annotation is available. Therefore, we augment this dataset with manual annotations of multiple facial attributes (see Table 7 for details). The annotated attributes[1] include: color of hair, visibility of eyes, type of eye wear, facial hair, whether mouth is open, smiling or not, wearing a hat, visibility of forehead, and gender. The annotations for this dataset will be made publicly available to the research community.[2]

The attributes were chosen to be objectively as unambiguous as possible while spanning a range of semantic properties with a variety of causal relationships amongst them as shown in Figure 1 of main paper. For example, smiling could be a cause of mouth being open because smiling might result in an open mouth. Or, wearing a hat could be a cause for affecting the visibility of forehead, since hats may cause occlusions on people's forehead. Moreover, gender could also causally affect facial hair, because females do not have facial hair in most cases.

8. Gradient of Closed-Form Solution

In order to find the gradient of the kernel ridge regressor of adversary, we rewrite the loss function of adversary as:

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

Then from [2], letting θ be arbitrary scalar element of encoder, we have

$$\frac{1}{2} \frac{\partial \|P_{\boldsymbol{K}} \boldsymbol{y}_f\|^2}{\partial \theta} = \boldsymbol{y}_f^T P_{\boldsymbol{K}^\perp} \frac{\partial \boldsymbol{K}}{\partial \theta} \boldsymbol{K}^\dagger \boldsymbol{y}_f, \tag{2}$$

where $\boldsymbol{K}^\perp$ is the orthogonal complement of $\boldsymbol{K}$, and

$$\left[\frac{\partial \boldsymbol{K}}{\partial \theta}\right]_{ij} = \begin{cases} \nabla_{\boldsymbol{z}_i}^T \left([\boldsymbol{K}]_{ij}\right) \nabla_\theta(\boldsymbol{z}_i) + \nabla_{\boldsymbol{z}_j}^T \left([\boldsymbol{K}]_{ij}\right) \nabla_\theta(\boldsymbol{z}_j), & i \leq n \\ 0, & \text{else.} \end{cases} \tag{3}$$

Equation 2 can be directly used to obtain the gradient of objective function in 1. The gradient of ridge regressor from unsupervised branch can be derived in same way by simply replacing the kernel matrix $\boldsymbol{K}$ with linear one.

[1] 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 task.

[2] The onus of obtaining the actual images will still remain with the respective research groups.

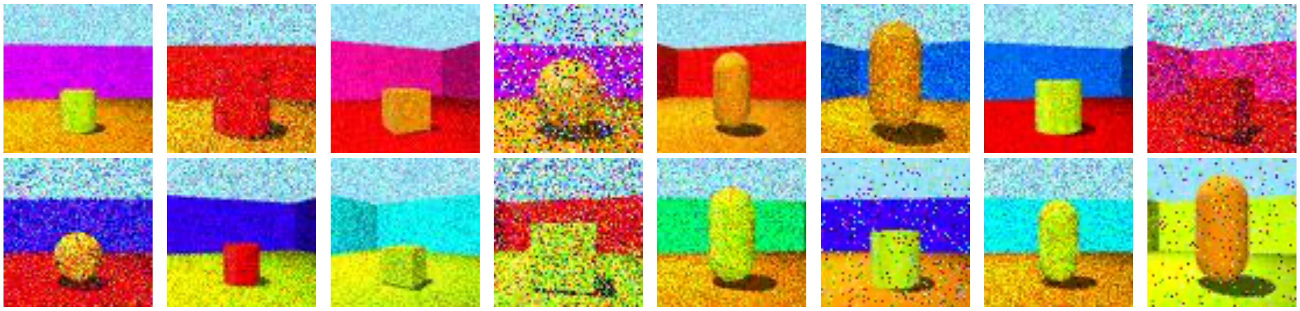

Figure 4. Sample images generated from the 3D Shapes dataset with known causal relations.

Table 7. CASIA-WebFace facial attributes, corresponding categories, and sample statistics.

Color of Hair		Eyes		Eye Wear		Facial hair		Forehead		Mouth		Smiling		Wearing a hat		Gender	
red	12,337	closed	18,047	none	424,128	none	364,076	partially visible	126,219	open	215,556	no	221,170	no	424,659	female	209,402
gray	17,050	open	425,185	eyeglasses	17,805	beard	1,763	visible	297,555	wide open	16,717	yes	231,890	yes	28,401	male	243,658
bald	13,239	not visible	9,828	sunglasses	11,127	mustache	21,525	fully blocked	29,286	closed	220,787						
blonde	85,848					goatee	2,613										
black	158,761					beard and mustache	48,025										
brown	144,523					mustache and goatee	15,058										
not visible	21,302																

References

[1] 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. 3

[2] Gene H Golub and Victor Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on numerical analysis*, 10(2):413–432, 1973. 4

[3] 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. 3

[4] Diviyan Kalainathan and Olivier Goudet. Causal discovery toolbox: Uncover causal relationships in python. *arXiv preprint arXiv:1903.02278*, 2019. 3

[5] Murat Kocaoglu, Alexandros G Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017. 1

[6] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision (ICCV)*, 2015. 4

[7] 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. 3

[8] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv:1411.7923*, 2014. 4