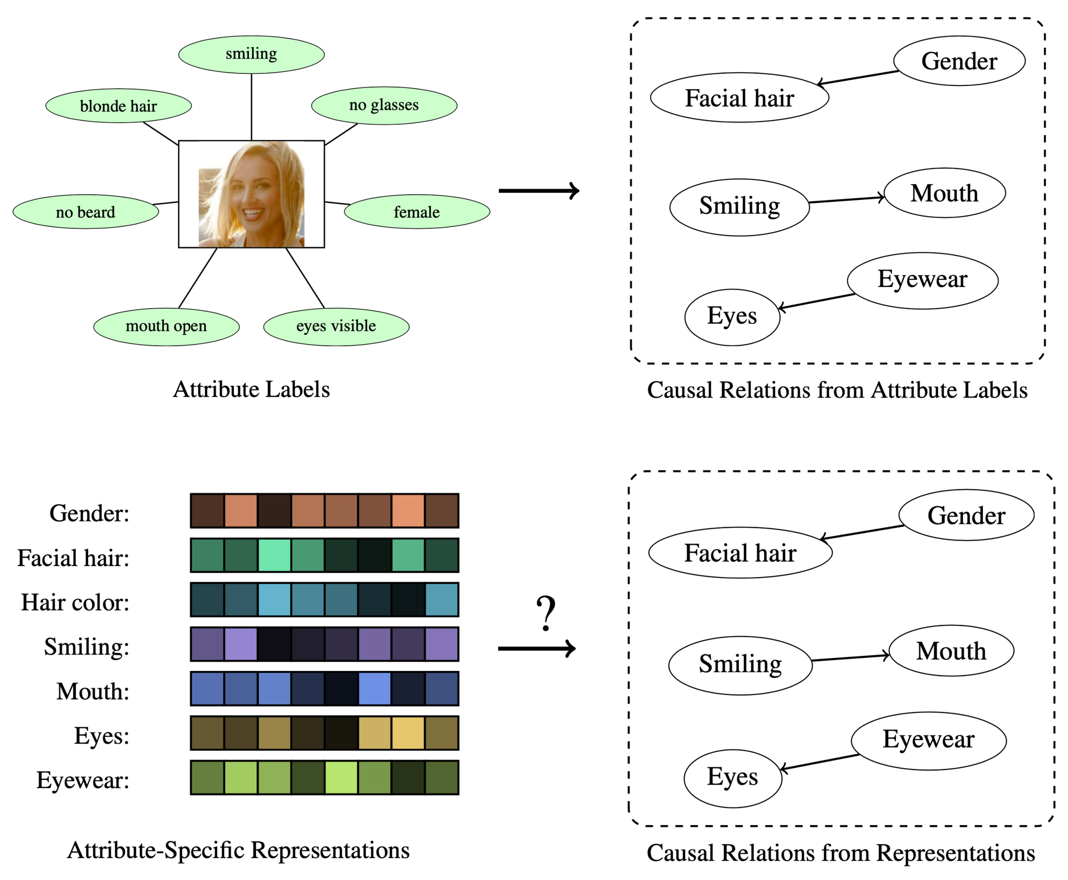
Do learned representations respect causal relationships?

Lan Wang and Vishnu Boddeti
Michigan State University

Motivation

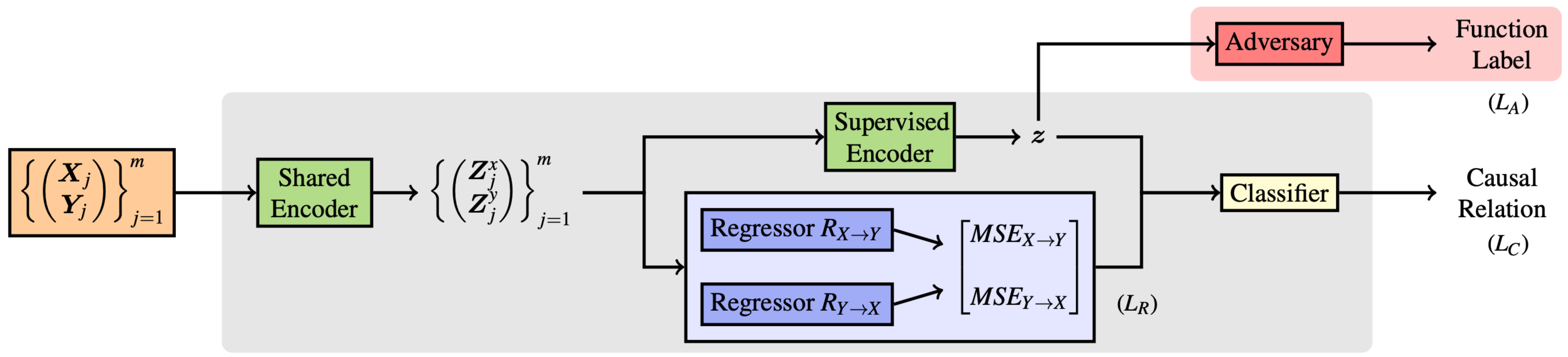


- Visual data may have multiple causally associated attributes.
- Do attribute-specific learned representations respect the underlying causal relationships between the attributes?

Contribution

- Propose Neural Causal Inference Net (NCINet) for causal discovery from high-dimensional observational data.
- Develop an experimental protocol
  - controllably resampled existing datasets to induce a known causal relation.
  - learn attribute representations and infer the causal relations.
- Analyze the effect on causal relation induced by various design in representation learning.

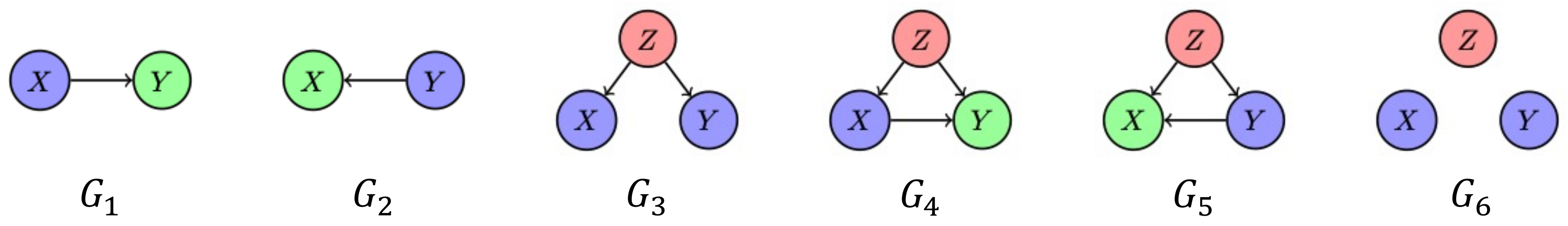
Proposed Method



Synthetic Causal Feature Generator | Unsupervised Branch | Adversarial Representation Learning | NCINet Inference

$$L_R = \frac{1}{m}\sum_{i=1}^{m}\|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2^2 + \frac{1}{m}\sum_{i=1}^{m}\|\hat{\boldsymbol{x}}_i - \boldsymbol{x}_i\|_2^2 \qquad L_A = -\|\boldsymbol{y}_f - \hat{\boldsymbol{y}}_f\|_2^2 = -\|\boldsymbol{y}_f - \boldsymbol{K}\left(\boldsymbol{K}+\beta\boldsymbol{I}\right)^{-1}\boldsymbol{y}_f\|_2^2 \qquad Loss = L_C + L_R + \lambda L_A$$

Causal Relations



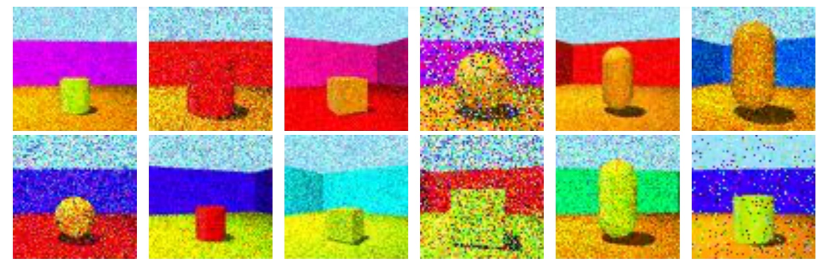
G_1	G_2	G_3	G_4	G_5	G_6
label=1	label=2	label=0	label=1	label=2	label=0

All possible causal relations between pairs of random variables.

Causal functions: Linear, Hadamard, Bilinear, Cubic Spline, Neural Network

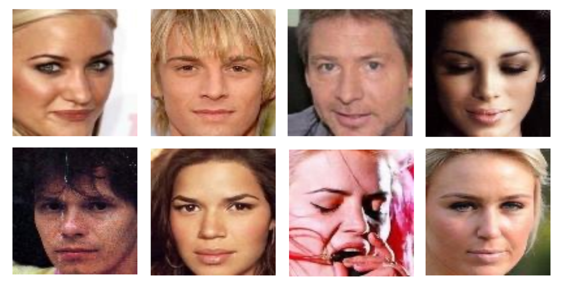
Dataset

Generate Labels: Generate labels of 6 graphs using Gibbs sampling (3 classes for both X, Y and Z)



Generate images: 3D shape

- Two factors are decided by X and Y: floor hue, wall hue.
- Exogenous variables: object hue, scale, shape, and orientation.
- Add random noise: Gaussian, Shot, or Impulse.



Sample images: CASIA-WebFace

- Annotations: color of hair, eyes, eye wear, facial hair, forehead, mouth, smiling, gender.
- Sample images with attributes consistent with generated labels.

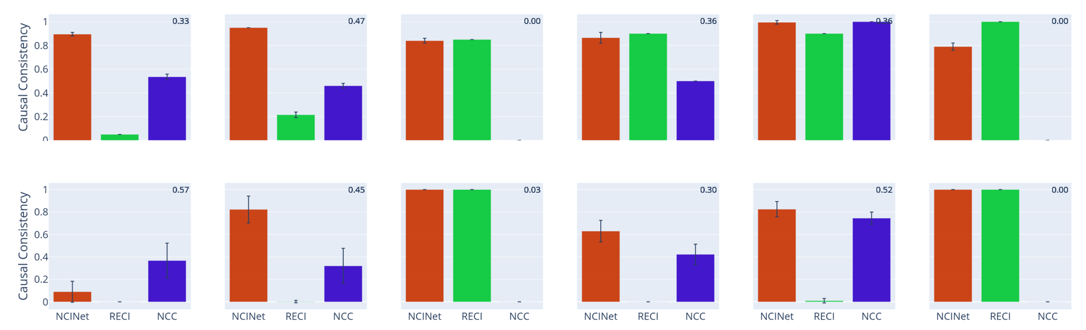
Experiments

Causal Consistency: $\frac{1}{K}\sum_{k=1}^{K}\frac{\#\text{consistent subsets}}{\#\text{subsets}}$

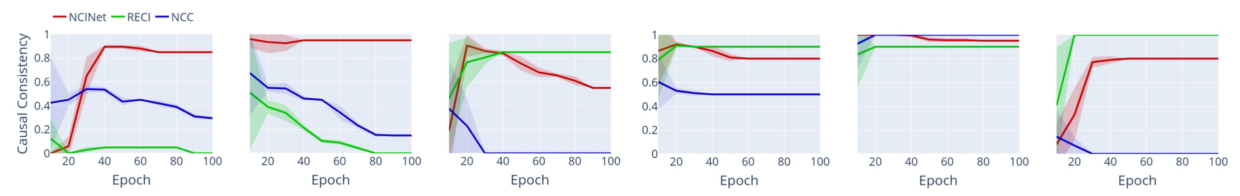
- Generalization Results

Methods	Linear	Hadamard	Bilinear	Cubic Spline	NN	Average
ANM [21]	31.87	32.49	32.94	33.66	33.08	32.81
Bfit [24]	34.89	54.76	53.69	**77.79**	38.26	51.88
NCC [37]	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**

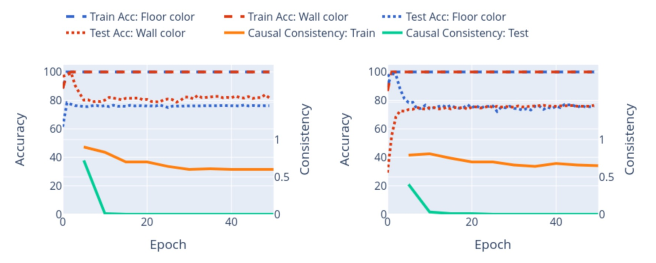
- Causal consistency on image datasets



- Effect of training epochs



- Effect of overfitting



Conclusion

- NCINet exhibits better causal inference generalization performance.
- Learned attribute-specific representations satisfy same causal relations attribute labels under controlled scenarios.
- More work is needed to force representations to learn causal relations.

Paper and Code