Measure-Theoretic Anti-Causal Representation Learning

Arman Behnam

Binghui Wang

Computer Science Department, Illinois Institute of Technology

Causality and Causal Representation Learning: Foundations

What is Causality?

- **Causation**: X causes Y if changing X would change Y
- **Association**: X and Y occur together (correlation)
- **Key distinction**: Causation implies *interventional* relationship

Causal Representation Learning:

- **Definition**: Study of discovering and leveraging causal relationships rather than mere statistical associations
- **Goal**: Identify *high-level causal variables* from low-level observations
- **Bridge**: Statistical pattern recognition $\rightarrow$ causal reasoning

Background: Causal Representation Learning Taxonomy

Distribution/Domain-Invariant Methods:

- **Classical**: IRM, REx, Domain Adaptation - seek invariant representations
- **Limitation**: No explicit causal modeling, assume I.I.D. data

Structure-Based Methods:

- **LECI, ICP, CSG**: Explicit DAG structure requirements
- **Limitation**: Require known SCM structure, limited to perfect interventions

Intervention-Based Methods:

- **ICRL, CIRL, iCaRL**: Model causal effects through interventions
- **Limitation**: Assume perfect interventions, need explicit SCM

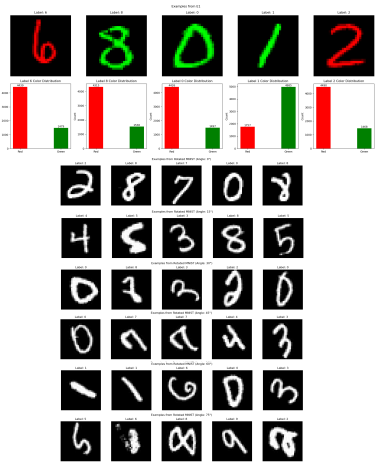
Related Works

Anti-Causal Representation Learning:

- *Theoretical*: Schölkopf et al. (2012) causal vs. anticausal paradigm
- *Recent advances*: Invariant representations (Jiang & Veitch, 2022)
- *Theory*: Schölkopf et al. (2021), Bengio et al. (2019) meta-transfer objectives
- *Interventional*: Ahuja et al. (2023), von Kügelgen et al. (2023)

Motivation

- Traditional ML assumes **causal direction**: features cause labels
- **Anti-causal** real problems: labels cause features
  - Disease causes symptoms (medical diagnosis)
  - Digit identity causes visual patterns
  - Tumor presence causes tissue appearance
- **Challenge**: Environment factors create spurious correlations
  - Hospital protocols affect medical images
  - Different imaging conditions change patterns
- **Goal**: Learn generalizable representations across environments by capturing true anti-causal mechanism



Problem

Anti-causal structure: $Y \to X \leftarrow E$

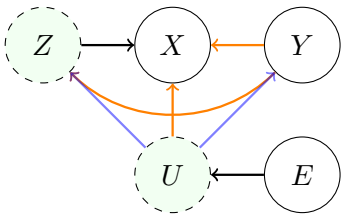
- Y: Label (disease, digit)
- X: Observations (symptoms, images)
- E: Environment (hospital, imaging setup)

Key Challenge:

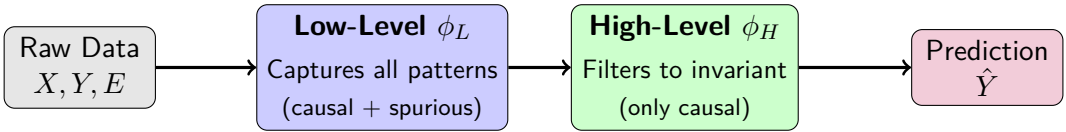
- Learn $Y \to X$ (true causal mechanism)
- Remove $E \to X$ (spurious correlation)
- Generalize to unseen environments

Existing Methods Limitations:

- Just for perfect interventions
- Need explicit DAG structure



ACIA: Two-Level Learning



Analogy: Recording audio

- ϕ_L: Microphone captures everything (voice + background noise)
- ϕ_H: Noise cancellation filters to just voice

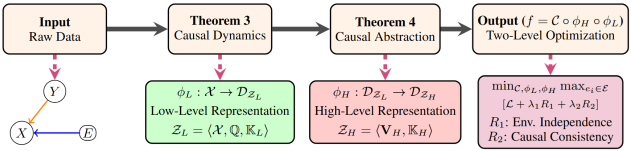
Key Takeaway

Why two levels? Single-level can't balance detail vs. invariance

Method

ACIA: Anti-Causal Invariant Abstractions
Key Innovation: Hierarchical representation

- **Low-level** ϕ_L: Captures raw anti-causal relationships
  - Extracts features from $Y \to X$ and $E \to X$
  - Preserves environment-specific information
- **High-level** ϕ_H: Distills invariant features
  - Removes environment effects
  - Retains only Y-relevant information



Measure-Theoretic Foundation: Why It Matters

Traditional Causality: Structural equations + DAGs

- Requires explicit causal graph
- Limited to perfect interventions

Measure-Theoretic Causality:

$$\text{Causal Space} = (\Omega_e, \mathscr{H}_e, P_e, K_e)$$

- Ω_e: Sample space (all possible states)
- $\mathscr{H}_e$: Event space (σ-algebra)
- P_e: Probability measure
- K_e: Causal kernel (the key!)

Advantage

Kernels K_e encode causal relationships *implicitly*[0.2cm] $\Rightarrow$ No DAG needed

Mathematical Foundations: Interventional Kernels

Key Theorem: Interventional Kernel Construction

$$K_S^{do(\mathcal{X},\mathbb{Q})}(\omega, A) = \int_\Omega K_S(\omega, d\omega')\mathbb{Q}(A|\omega') \tag{1}$$

Perfect vs Imperfect Interventions:

- **Perfect**: $\mathbb{Q}(A|\omega') = \mathbb{Q}(A)$ (independent of ω')
- **Imperfect**: $\mathbb{Q}(A|\omega') = \alpha \cdot P(A|Y = y') + (1 - \alpha) \cdot P(A)$

Anti-causal Property:

- $K_S^{do(X)}(\omega, \{Y \in B\}) = K_S(\omega, \{Y \in B\})$
- $K_S^{do(Y)}(\omega, \{X \in A\}) \neq K_S(\omega, \{X \in A\})$

Regularizer Implementation Details

Environment Independence (R_1):

$$R_1 = \sum_{e_i, e_j \in \mathcal{E}, i \neq j} \left\| \mathbb{E}[\phi_H(\phi_L(X)) | Y = y, E = e_i] - \mathbb{E}[\phi_H(\phi_L(X)) | Y = y, E = e_j] \right\|_2 \quad (2)$$

Causal Structure Consistency (R_2):

$$R_2 = \sum_{e_i \in \mathcal{E}} \left\| \mathbb{E}[Y | \phi_H(\phi_L(X)), E = e_i] - \mathbb{E}[Y | K_{\{e_i\}}^{do(Y)}] \right\|_2 \quad (3)$$

Practical Implementation:

- R_1: Minimize KL divergence between conditional distributions
- R_2: Align predictions with interventional distributions
- $\lambda_1, \lambda_2 = O(1/\sqrt{n})$ for convergence guarantees

Theoretical Guarantees

OOD Generalization Bound:

$$\mathbb{E}_{e_{\text{test}}}[\ell(f^*)] \leq \max_{e \in \mathcal{E}} \mathbb{E}_e[\ell(f^*)] + O\left(\sqrt{\frac{\log(1/\delta)}{n_{test}}}\right) \tag{4}$$

Performance Gap Lower Bound:

$$|\mathbb{E}_{e_{\text{test}}}[\ell(f^*)] - \mathbb{E}_{e_{\text{train}}}[\ell(f^*)]| \geq \min_{e \in \mathcal{E}} \|K_{\{e\}}^{do(X)} - K_{\{e\}}\|_{\mathcal{H}} \tag{5}$$

Convergence Rate:

- Distance to optimum: $O(1/\sqrt{T}) + O(1/\sqrt{n})$
- T: iterations, n: sample size

Key Takeaway

First anti-causal method with provable OOD generalization for imperfect interventions

Experiments & Results

Datasets: CMNIST, RMNIST, Ball Agent, Camelyon17

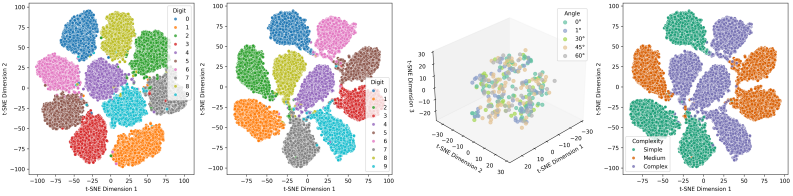
Dataset	Test Acc	Env. Indep.	Low-level Inv.	Interv. Robust.
CMNIST	**99.2%**	**0.00**	**0.01**	**0.02**
Best Baseline	95.5%	0.64	0.40	0.67
RMNIST	**99.1%**	**0.00**	**0.03**	**0.01**
Best Baseline	93.5%	0.23	2.77	0.04
Ball Agent	**99.98%**	0.52	**0.03**	**0.03**
Best Baseline	74.0%	0.46	0.39	0.05
Camelyon17	**84.4%**	0.28	**0.42**	**0.43**
Best Baseline	65.5%	0.23	0.50	0.45

Key Achievements:

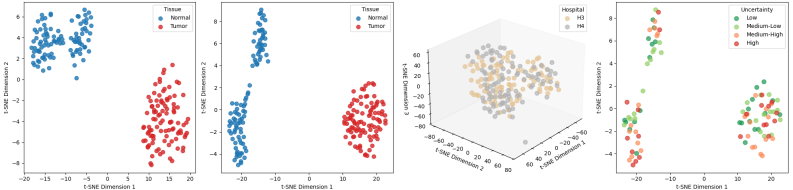
- **Perfect environment independence** on synthetic datasets
- **Robust to both perfect and imperfect interventions**

Experiments & Results - Visualization of Learned Representations

RMNIST Results - Two-Level Disentanglement



Camelyon17 Results - Medical Data Generalization



Technical Challenges and Solutions

1. **Challenge**: Single-level representations can't balance causal preservation vs environment invariance
   Solution: Two-level hierarchy
   - ϕ_L preserves rich anti-causal structure
   - ϕ_H abstracts to environment-invariant features

2. **Challenge**: Handling imperfect interventions in real-world scenarios
   Solution: Interventional kernel $K_S^{do(\mathcal{X},\mathbb{Q})}(\omega, A) = \int K_S(\omega, d\omega')\mathbb{Q}(A|\omega')$
   - Perfect interventions: $\mathbb{Q}(A|\omega') = \mathbb{Q}(A)$
   - Imperfect interventions: $\mathbb{Q}(A|\omega')$ depends on ω'

3. **Challenge**: Learning without explicit DAG requirements
   Solution: Measure-theoretic causality (Product Causal Space & Sub-σ-algebra)
   - Direct learning from anti-causal structure
   - Theoretical OOD generalization guarantees

ACIA vs. State-of-the-Art: Comprehensive Comparison

Method	Anti-causal	No SCM	Imperfect Int.	Nonparam.	OOD
Distribution-Invariant Methods					
IRM, REx, Domain Adapt.	✗	✓	✗	✗	✓
CausalDA, Transportable Rep	✓	✓	✗	✓	✓
Structure-Based Methods					
ICP, CSG, LECI	✗	✗	✓	✓	✓
Causal Disentanglement	✗	✗	✗	✗	✓
Intervention-Based Methods					
ICRL, CIRL, iCaRL	✗	✗	✓	✓	✓
Nonparametric ICR	✗	✗	✓	✓	✓
ACIA (Ours)	✓	✓	✓	✓	✓

Key Takeaway

ACIA is the **only** method satisfying all five critical properties simultaneously

Contributions

Theoretical:

- For perfect/imperfect interventions through interventional kernels
- Causal features identification across environments without requiring explicit DAG specifications
- Out-of-Distribution (OOD) generalization bounds

Practical:

- Capture and exploit anti-causal structures in synthetic and real-world datasets
- Superiority over state-of-the-arts