Modeling the Data-Generating Process is Necessary for Out-of-Distribution Generalization

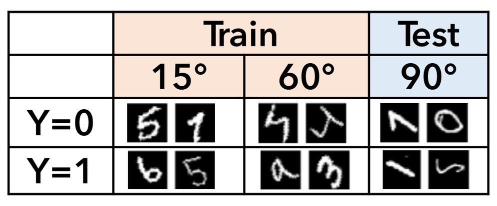
Jivat Neet Kaur   Emre Kıcıman   Amit Sharma

Microsoft

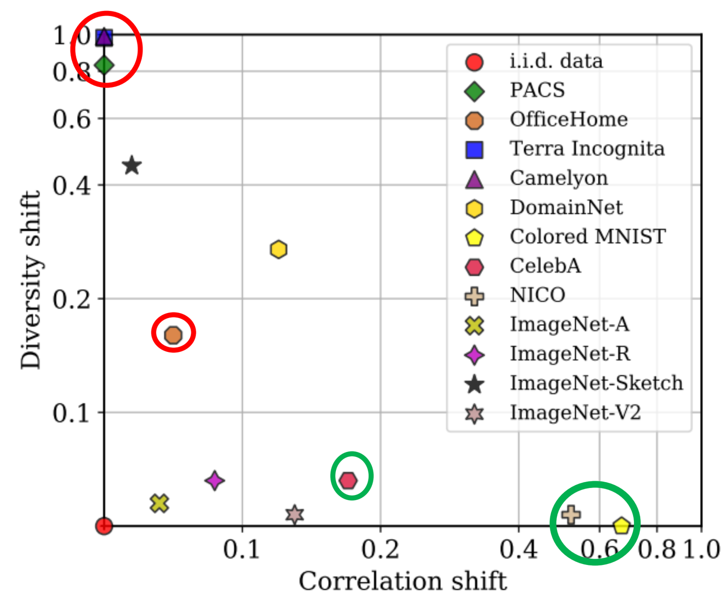
{t-kaurjivat, emrek, amshar} @ microsoft.com

https://arxiv.org/abs/2206.07837

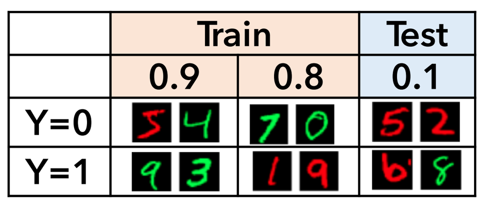
State of SoTA Domain Generalization Algorithms



Rotated MNIST



Algorithm	Ranking score
MMD [42]	+2
ERM [69]	0
VREx [38]	-1
GroupDRO [6	-1



Colored MNIST

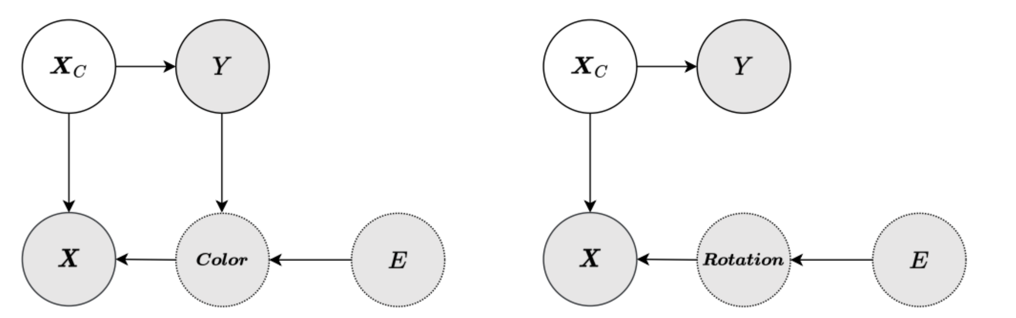
Algorithm	Ranking score
VREx [38]	+1
GroupDRO [63]	+1
ERM [69]	0
MMD [42]	-1

No method can surpass ERM on all kinds of shifts!

[1] Ye et al., CVPR 2022

Distribution Shifts: Causal Perspective

- Different distribution shifts arise due to **differences in data-generating process (DGP)**
  - Leading to different independence constraints

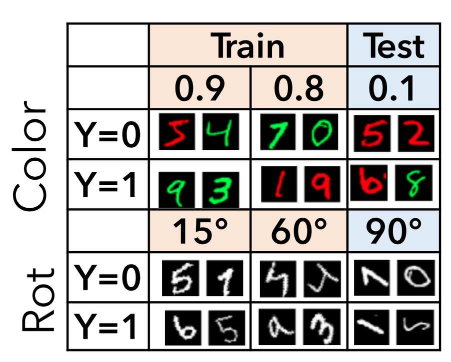


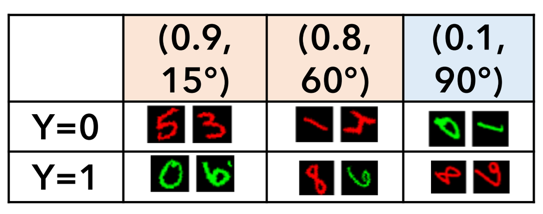
- Any algorithm based on a single, fixed independence constraint cannot work well across all shifts

Solution: Modeling the causal relationships in DGP

Multi-attribute Distribution Shifts

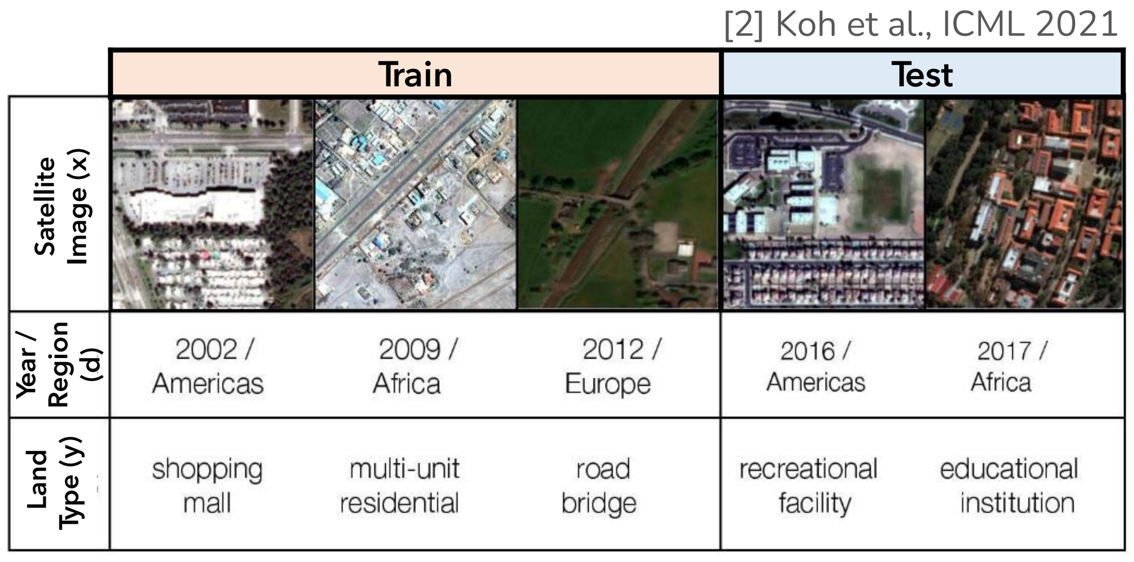
What if different distribution shifts co-exist?





Col+Rot

[2] Koh et al., ICML 2021

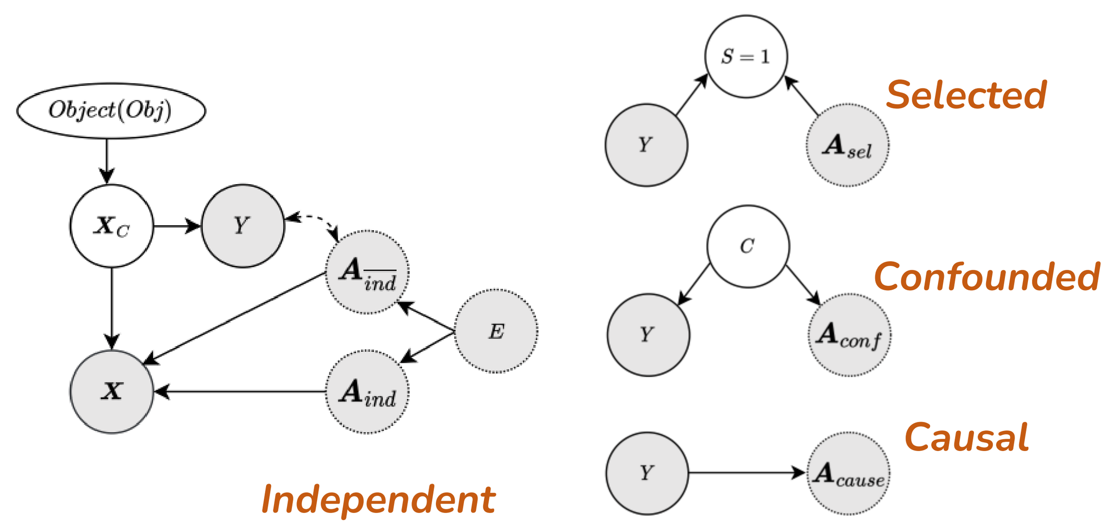


Real-world data contains shifts on multiple attributes

Can we develop an algorithm that generalizes to not just individual shifts, but also multi-attribute shifts?

Causally Adaptive Constraint Minimization (CACM)

Generalization under *Independent*, *Causal*, *Confounded* and *Selected* shifts



Causal DAG to specify multi-attribute shifts

Different $Y - A_{\overline{ind}}$ relationships

Theorem.

1. *Independent*: $X_c \perp\!\!\!\perp A_{ind}$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{ind} | Y$; $X_c \perp\!\!\!\perp A_{ind} | E$; $X_c \perp\!\!\!\perp A_{ind} | Y, E$
2. *Causal*: $X_c \perp\!\!\!\perp A_{cause} | Y$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{cause} | Y, E$
3. *Confounded*: $X_c \perp\!\!\!\perp A_{conf}$; $X_c \perp\!\!\!\perp E$; $X_c \perp\!\!\!\perp A_{conf} | E$
4. *Selected*: $X_c \perp\!\!\!\perp A_{sel} | Y$; $X_c \perp\!\!\!\perp A_{sel} | Y, E$

Observation: Note that no constraint is valid across all four settings

Theorem. For any predictor algorithm for Y that uses a single type of (conditional) independence constraint, there exists a realized graph $\mathcal{G}$ and a corresponding training dataset such that the learned predictor cannot be a risk-invariant predictor across distributions in $\mathcal{P}_{\mathcal{G}}$.

Therefore, we propose an algorithm that adaptively applies the right constraint.

Algorithm for general graph

Phase I: Derive correct independence constraints

1. For every observed variable $A \in \mathcal{A}$ in the graph, check whether (X_c, A) are d-separated.
   => $X_c \perp\!\!\!\perp A$ is a valid constraint
2. If not, check whether (X_c, A) are d-separated conditioned on any subset A_s of the remaining observed variables in $\mathcal{A} \setminus \{A\}$.
   => $X_c \perp\!\!\!\perp A | A_s$ is a valid constraint

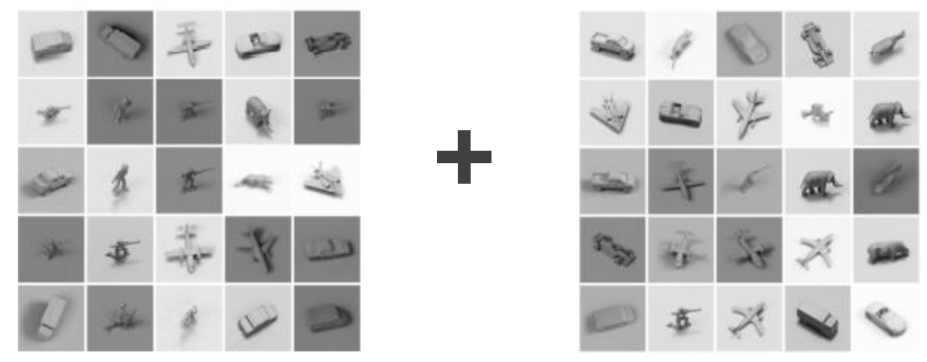
Phase II: Apply regularization penalty using constraints derived

$$RegPenalty = \sum_{A \in \mathcal{A}} Penalty_A$$

Algorithm	Color	Rotation	Col+Rot
ERM	30.9 ± 1.6	61.9 ± 0.5	25.2 ± 1.3
IRM	50.0 ± 0.1	61.2 ± 0.3	39.6 ± 6.7
MMD	29.7 ± 1.8	62.2 ± 0.5	24.1 ± 0.6
C-MMD	29.4 ± 0.2	62.3 ± 0.4	32.2 ± 7.0
CACM	**70.4 ± 0.5**	**62.4 ± 0.4**	**54.1 ± 0.3**

Empirical Evaluation

Correct constraint derived from causal graph matters



spurious correlation b/w category and lighting (A_{cause})

Unseen data shift: unseen azimuth values (A_{ind})

small NORB dataset

- Multi-class (5 classes)
- Muti-valued attributes
- Real objects

[3] Wiles et al., ICLR 2022

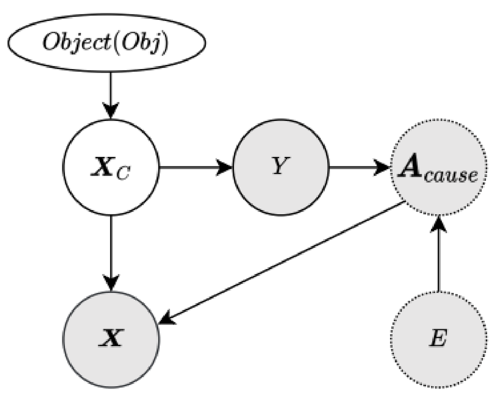
Algorithm	lighting A_{cause}	azimuth A_{ind}	lighting+azimuth $A_{cause} \cup A_{ind}$
ERM	65.5 ± 0.7	78.6 ± 0.7	64.0 ± 1.2
IRM	66.7 ± 1.5	75.7 ± 0.4	61.7 ± 1.5
VREx	64.7 ± 1.0	77.6 ± 0.5	62.5 ± 1.6
MMD	66.6 ± 1.6	76.7 ± 1.1	62.5 ± 0.3
CORAL	64.7 ± 1.5	77.2 ± 0.7	62.9 ± 0.3
DANN	64.6 ± 1.4	78.6 ± 0.7	60.8 ± 0.7
C-MMD	65.8 ± 0.8	76.9 ± 1.0	61.0 ± 0.9
CDANN	64.9 ± 0.5	77.3 ± 0.3	60.8 ± 0.9
CACM	**85.4 ± 0.5**	**80.5 ± 0.6**	**69.6 ± 1.6**

No single algorithm performs well across all shifts

CACM provides upto 20% improvement

Incorrect constraints hurt generalization!

I.



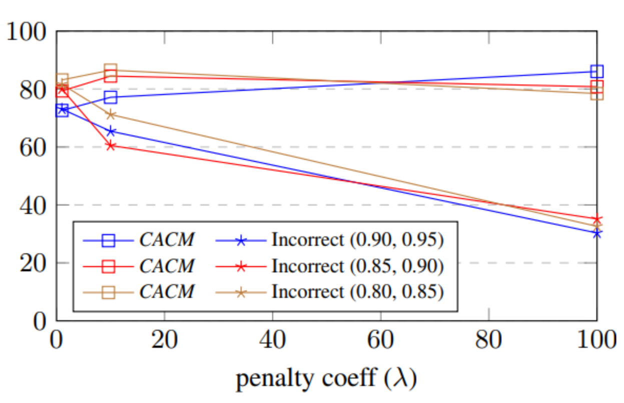
Causal

$X_c \perp\!\!\!\perp A_{cause} | E$ ✗

$X_c \perp\!\!\!\perp A_{cause} | Y, E$ ✓

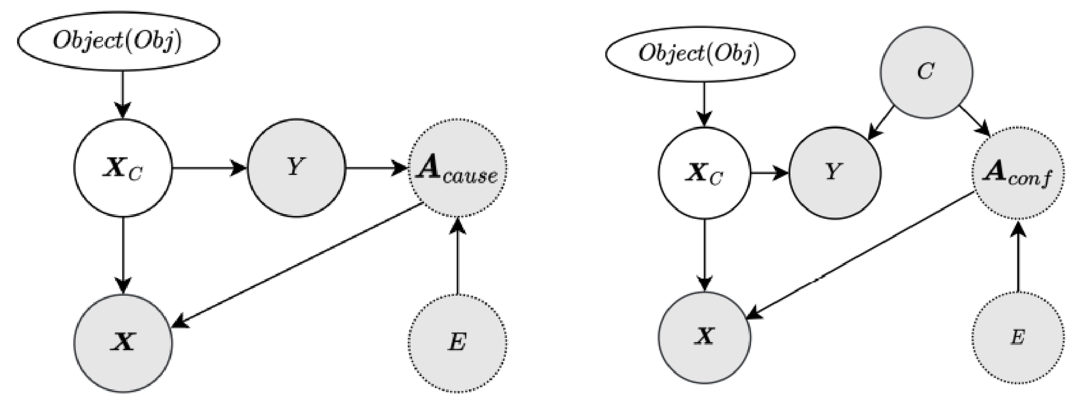
Constraint	Accuracy
$X_c \perp\!\!\!\perp A_{cause} \| E$	29.7 ± 3.8
$X_c \perp\!\!\!\perp A_{cause} \| Y, E$	**94.1 ± 0.5**

Synthetic dataset



small NORB dataset

II.



Causal

$X_c \perp\!\!\!\perp A_{cause} | E$ ✗

$X_c \perp\!\!\!\perp A_{cause} | Y, E$ ✓

Confounded

$X_c \perp\!\!\!\perp A_{conf} | E$ ✓

$X_c \perp\!\!\!\perp A_{conf} | Y, E$ ✗

Constraint	*Causal*	*Confounded*
$X_c \perp\!\!\!\perp A \| E$	29.7 ± 3.8	**62.4 ± 1.9**
$X_c \perp\!\!\!\perp A \| Y, E$	**94.1 ± 0.5**	56.0 ± 1.0

Conclusion

- Important to study *multi*-attribute shifts
- Algorithms based on single, fixed constraint fail
- Necessary to model causal relationships in the data-generating process