NEURAL INFORMATION PROCESSING SYSTEMS

LIVS @ UBCO

UBC

Causal Gradient Steering

Exposing Shortcut Gradients by Destroying Causal Signal

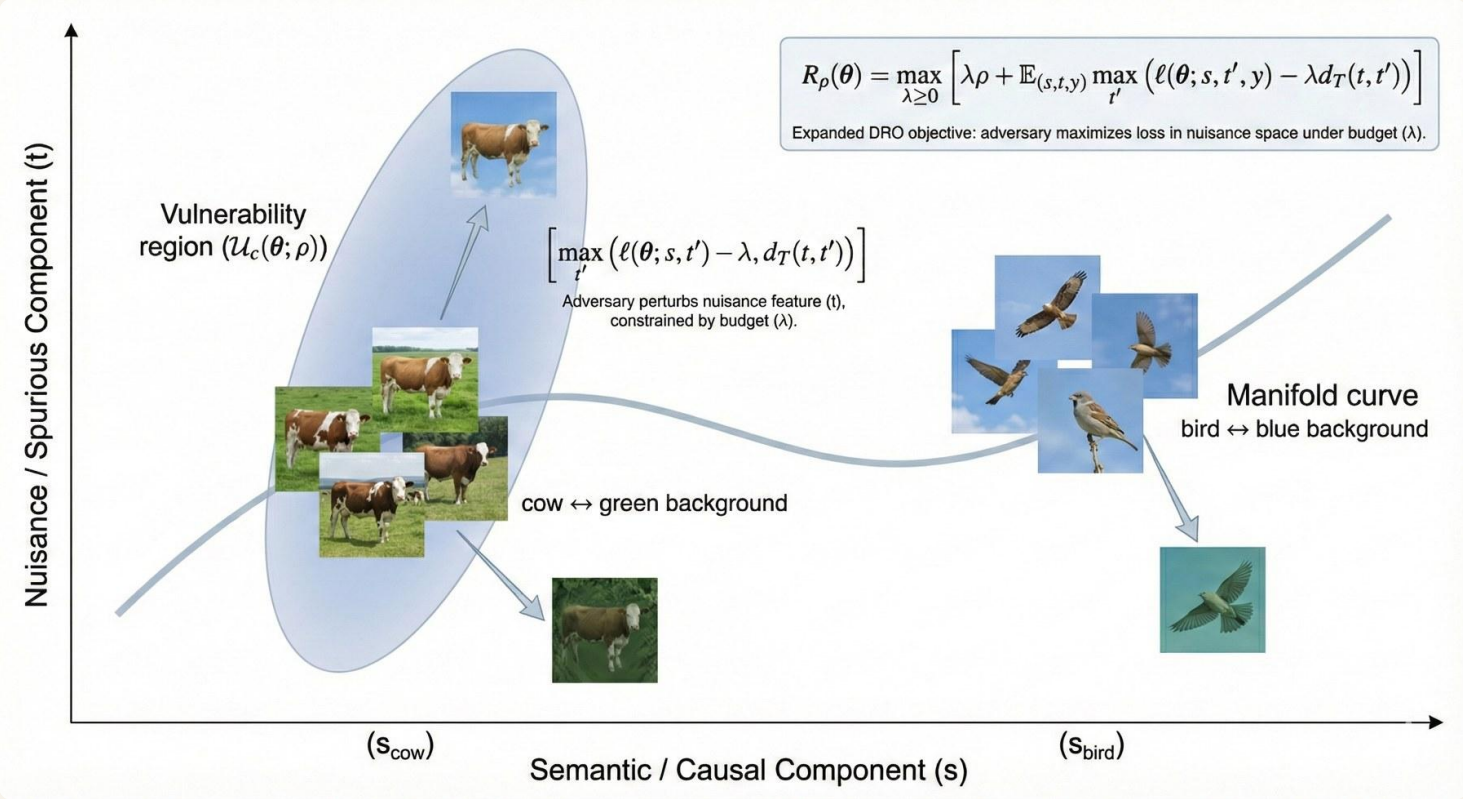
Ahmed Radwan, Ahmad Abdel-Qader, Mahmoud Soliman, Omar Abdelaziz, Ahmed Elgazwy, Mohamed Shehata, Shan Du

Department of Computer Science · The University of British Columbia · Kelowna, BC, Canada

PROBLEM

ERM cannot tell the cow from the grass

On biased data, both the **object** and its **background** lower the loss, so the model learns both.



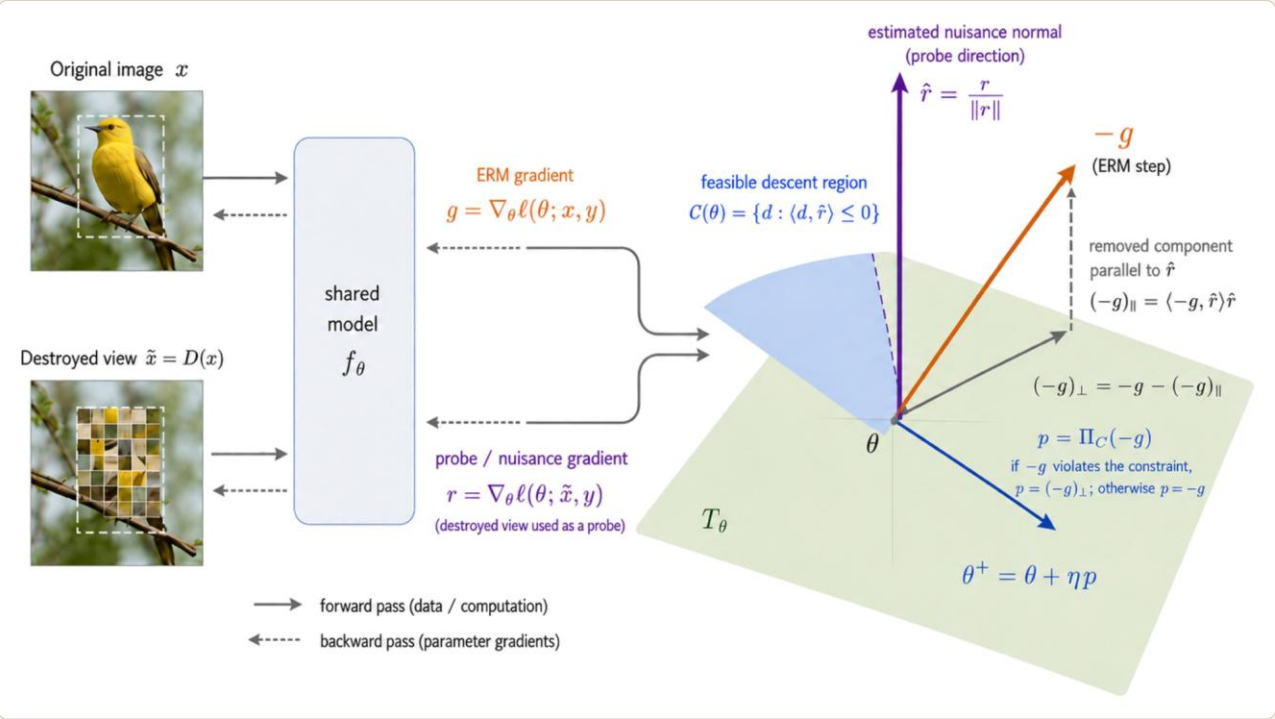
KEY IDEA

It is easier to destroy the object than to isolate it.

Shuffle patches inside the object box: shape is gone, colour, texture and background remain. If the model still predicts the label, that gradient points straight at the shortcut.

METHOD

Steer the update away from the shortcut



1 · Clean image → g

The usual ERM gradient.

2 · Destroyed view → r

Used only as a probe, never as a training sample.

3 · Project

If the ERM step agrees with r, remove its part along r. Otherwise keep the ERM step.

Closed-form projection · **Probe every 100 steps · ~1% cost** · **No domain labels** · **Zero inference cost**

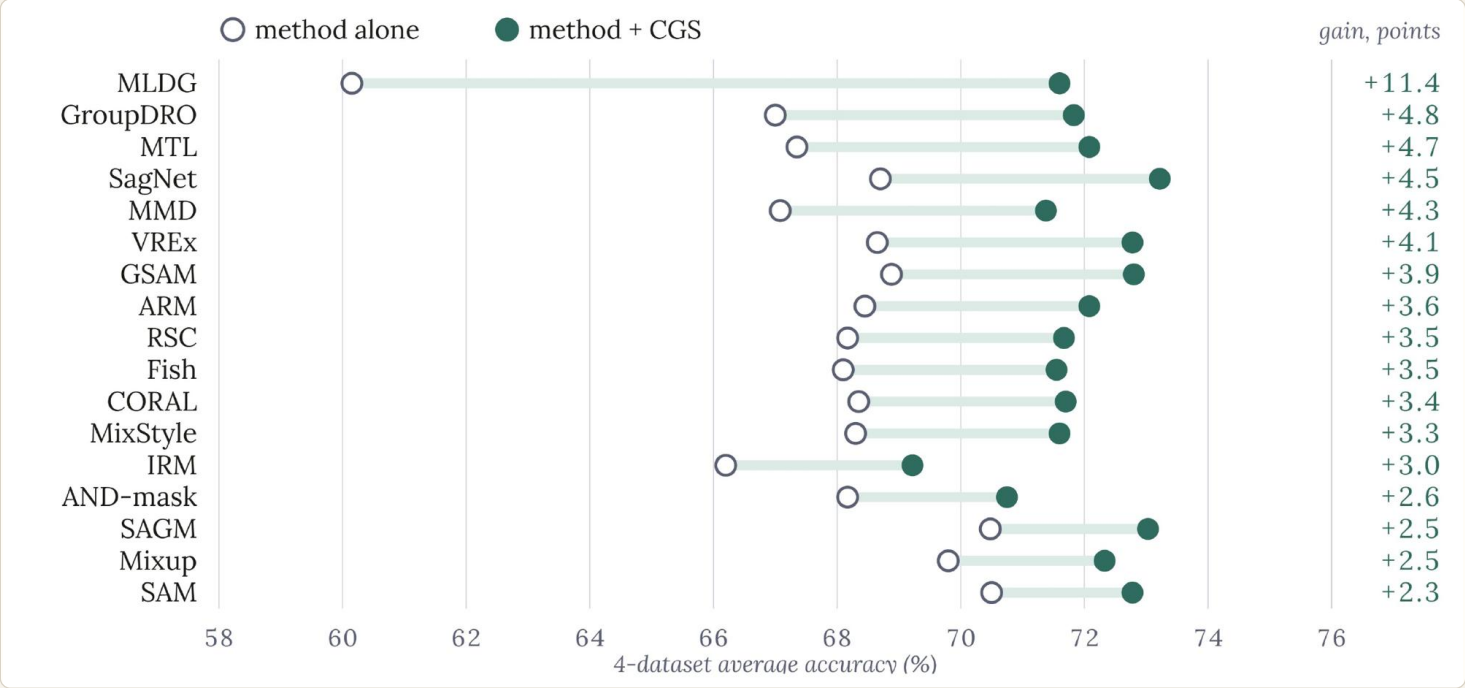
WHERE THE MODEL LOOKS

CGS attends to the object, not the context



PLUG-IN

Adding CGS improves every DG method we tried



17 / 17

methods improve, by **+2.3 to +11.4** points on average.

One locked CGS configuration, no per-method retuning.

BEYOND CLASSIFICATION

Design a destroyer for any task

Like an augmentation, but inverted: break the causal cue, keep everything else.

Classification

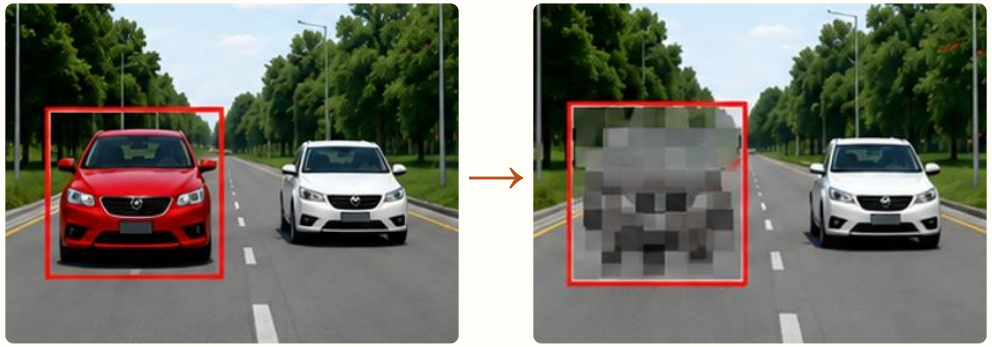


Destroy object shape

Exposes reliance on branch, leaves, colour

DomainBed avg 63.4 → 69.0 vs ERM

Tracking



Destroy the target in the search frame

Exposes locating by position and scene, not appearance

OSTrack-256 GOT-10k AO 71.0 → 72.4

CONCLUSION

Main points

- Destroying the object in an image reveals the gradient that comes from shortcuts.
- CGS removes that gradient from each update, at about 1% extra training cost.
- Destroyers can be designed for each task, much like data augmentations.

github.com/AhmedMostafaSoliman/CausalGradientSteering

LINKEDIN

Thank you.