GRIT · GEOMETRIC ROBUSTNESS VIA INVARIANT TRAINING

Provable Robustness to Spurious Correlations via Invariant Data

Ruqi Bai[1] Yao Ji[2] Easton Currie[1] Mingyu Kim[1] Zeyu Zhou[1] David I. Inouye[1] [1]Purdue University [2]Georgia Institute of Technology

PURDUE · GEORGIA TECH
Midwest Machine Learning Symposium
MMLS 2026 · West Lafayette, IN

◆ THE TAKE-HOME MESSAGE

No causal graph. No new architecture. Just the right data.

We decouple OOD robustness into a *data problem* and an *algorithm problem*— then prove our approach is robust even with a small number of (noisy) invariant pairs.

THE DATA PROBLEM
Invariant Data Pairs
application-specific — same content, different spurious content

\+

THE ALGORITHM PROBLEM
Orthogonality Constraint
application-agnostic — project the classifier off the spurious subspace

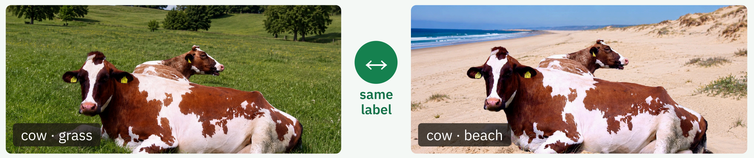
=

PROVABLE — EVEN WITH FEW NOISY PAIRS
Provable OOD Robustness
spurious error decays as $1/\sqrt{k}$ in the number of pairs

01 Invariant Data Pairs

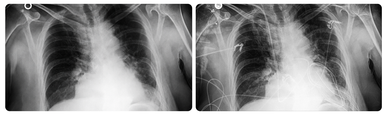
"Same semantic content—different spurious content."

◆ ONE INVARIANT DATA PAIR



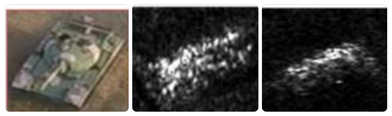
The animal and its label are unchanged — only the **spurious background** changes from grass to beach.

MORE INVARIANT PAIRS — ACROSS DOMAINS

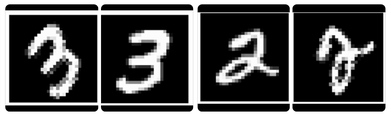


Chest X-ray — same patient, **with vs. without medical devices**

Aerial / satellite — same land, **different weather**



SAR target (ATR) — same target, **view angle 37° vs. 31°**

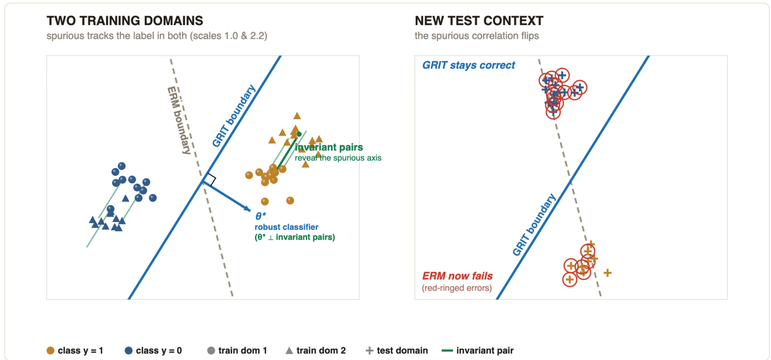


Handwritten digit — same digit & style, **different rotation**

Pairs may be **noisy, imperfect, even unlabeled** — and any pair of **different domains** works, since changing the environment implicitly changes the spurious content.

02 Orthogonality Constraint

The pair differences reveal the spurious subspace/direction — force the classifier to be orthogonal to these differences.



Two training domains share the **same** spurious correlation, so ERM's boundary leans on it — and fails when the correlation **flips** in a new context (**red errors, right**). The **invariant pairs** expose the spurious direction; GRIT holds the classifier **θ* orthogonal to that direction**, so it relies on the stable core alone.

$$\min_{\theta} \mathbb{E}_{\text{train}}\big[\ell(\theta^\top \phi(x),\, y)\big] \quad \text{s.t.} \quad \theta^\top \tilde{Q}_r = 0$$

$$\tilde{Q}_r = \text{top-}r \text{ eigenvectors of } \tilde{\Sigma}_k = \frac{1}{k}\sum_{i=1}^{k} \delta_i \delta_i^\top$$

- A standard ERM solve **plus a linear constraint**.
- Application-agnostic — no causal model, no architecture change.
- Even a **single clean pair** gives a useful direction.

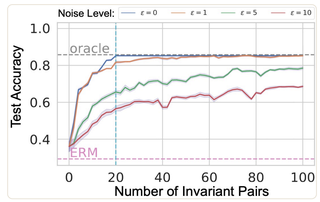
03 Provable Robustness

Bounded spurious error — shrinking as you add pairs.

$$\text{Spurious error} \le \text{Test shift} \times \frac{\text{pair noise}}{\text{pair signal}} \times \frac{1}{\sqrt{k}}$$

Test shift — distance to the new context. **Pair SNR** — strong spurious change + tight match.
1/√k — only a few pairs are needed.

RESULTS



SYNTHETIC VALIDATION

Test accuracy climbs to the *oracle* as pairs *k* grow — matching the O(1/√k) rate and robust to pair noise ε. Dashed: ERM.

EMPIRICAL — FROZEN-CLIP LINEAR PROBE

METHOD	C-MNIST test acc	WATERBIRDS worst-group
ERM (CLIP)	9.3	78.1
GroupDRO	12.7	68.4
Fish	11.8	74.4
MatchDG	18.1	53.6
GRIT (ours)	**69.3**	**81.2**

Worst-group / test accuracy (%). GRIT is best on both — *+51 pts on ColoredMNIST, +3 pts on Waterbirds over the strongest baseline.*

PROBLEM SETUP

We study **robust finetuning with a fixed representation** $z = \phi(x) \in \mathbb{R}^d$; the two assumptions below are stated informally.

Assumption 1 (informal) — Stable subspace. A linear subspace S on which the joint distribution of label and features is invariant across all environments e:

$$\mathbb{P}^{(e)}(y,\, z_{\mathcal{S}}) = \mathbb{P}^{(e')}(y,\, z_{\mathcal{S}}) \quad \forall\, e, e'$$

Assumption 2 (informal) — Noisy invariant pairs. Pair differences lie in the complement U = $S^\perp$ up to isotropic sub-Gaussian noise, and the signals span U:

$$\delta_i = \phi(x_i) - \phi(x_i') = u_i + \xi_i, \quad u_i \in \mathcal{U} := \mathcal{S}^\perp$$

Special cases: latent causal models (linear observation) · multi-view models (shared observation) · data augmentation (implicitly defines the invariant spaces).

FORMAL GUARANTEE

Under Assumptions 1–2, with rank $r \ge d_{sp}$ and $k \ge d$ pairs, with probability at least $1 - \eta$ the test-domain log-loss obeys:

$$\mathbb{E}_{\text{test}}[\ell] \le \mathbb{E}_{\text{train}}[\ell] + C\,\|\theta\| \sqrt{\lambda_1^{\text{test}}} \cdot \sqrt{\frac{\sigma_\xi^2 + \sigma_\xi \sqrt{\lambda_1}}{\lambda_{d_{sp}}}} \cdot \sqrt{\frac{d \log(2d/\eta)}{k}}$$

The three factors are exactly the **test-shift strength** λ_1^{test}, the **pair signal-to-noise** $(\sigma_\xi, \lambda_{d_{sp}})$, and the **1/√k** sample rate of the friendly bound above.

The first finite-sample analysis for this setting. Analogous bounds hold for regression with squared-error loss.

EXPERIMENTS

Benchmarks. ColoredMNIST & Waterbirds-CF, where color / background create the spurious correlation.

Backbone. Frozen CLIP embeddings; only the final linear classifier is trained.

Metric. Worst-group accuracy under in-domain validation, vs. ERM and DG baselines.

DISCUSSION — SOURCING PAIRS

Creating pairs is *not* our focus — but the decoupling turns it into a separate, tractable data problem. Three practical routes:

Measure twice (multi-view) — re-capture the same subject under a changed condition.

Pair existing data — match two training samples close in semantics but differing in a spurious attribute (e.g. rotated vs. non-rotated).

Generate pairs — prompt a foundation model to change only the spurious feature.

Acknowledgements NSF IIS-2212097 · ARL W911NF-2020-221 · ONR N00014-23-C-1016. Opinions, findings, and conclusions are the authors' own and do not necessarily reflect the sponsors'.