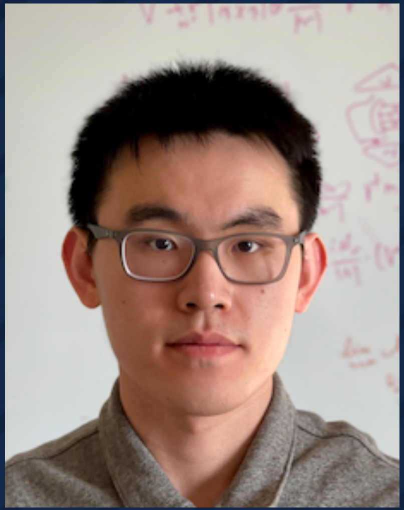
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Provable Domain Generalization via Invariant-Feature Subspace Recovery

ICML 2022



Haoxiang Wang
PhD Candidate
ECE, UIUC

Haozhe Si
Master Student
ECE, UIUC

Bo Li
Assistant Professor
Computer Science, UIUC

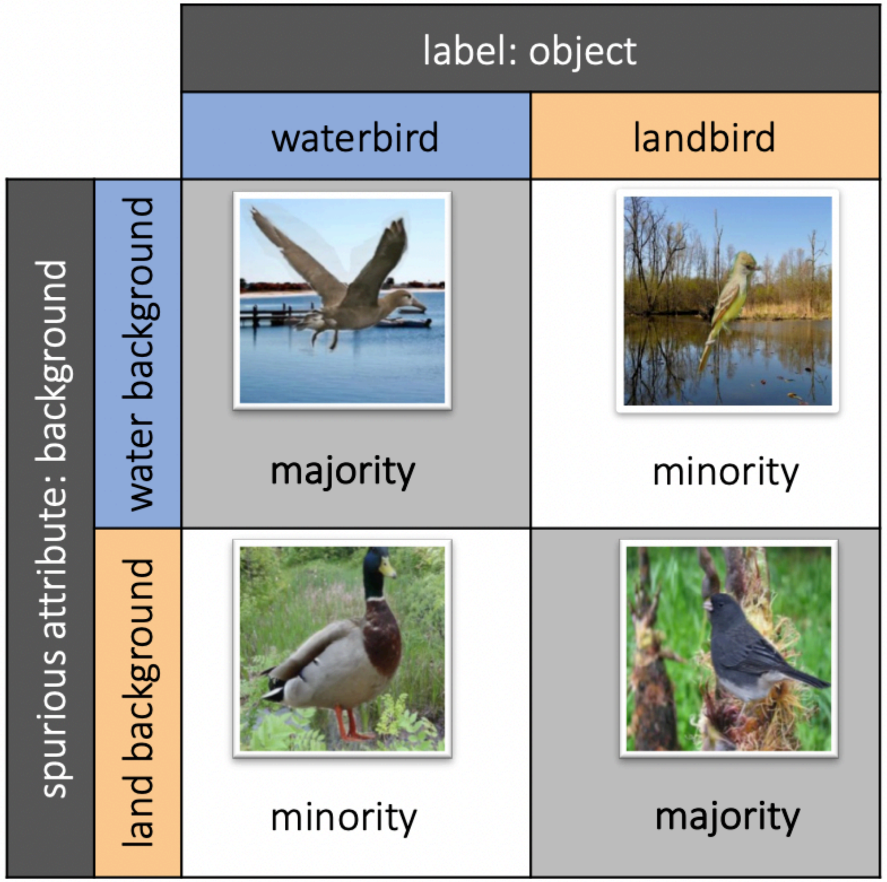
Han Zhao
Assistant Professor
Computer Science, UIUC

Task: Domain Generalization — Spurious Correlations

Domain Generalization
(OOD Generalization)

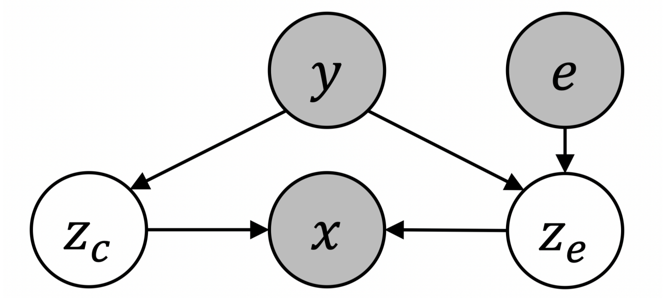
Perspective →

Spurious Correlation



Left Image: Li et al. *Deeper, Broader and Artier Domain Generalization*. ICCV 2017.   Right Image: Sagawa et al. *Distributionally Robust Neural Networks*. ICLR 2020.

Problem Setup and Data Generation Process [Rosenfeld at al. ICLR 2021]



In arbitrary training environment $e = 1,2,\ldots,E$, a sample (x, y, e) is generated by:

$$y = \begin{cases} 1, & \text{with probability } \eta \\ -1, & \text{otherwise} \end{cases}$$

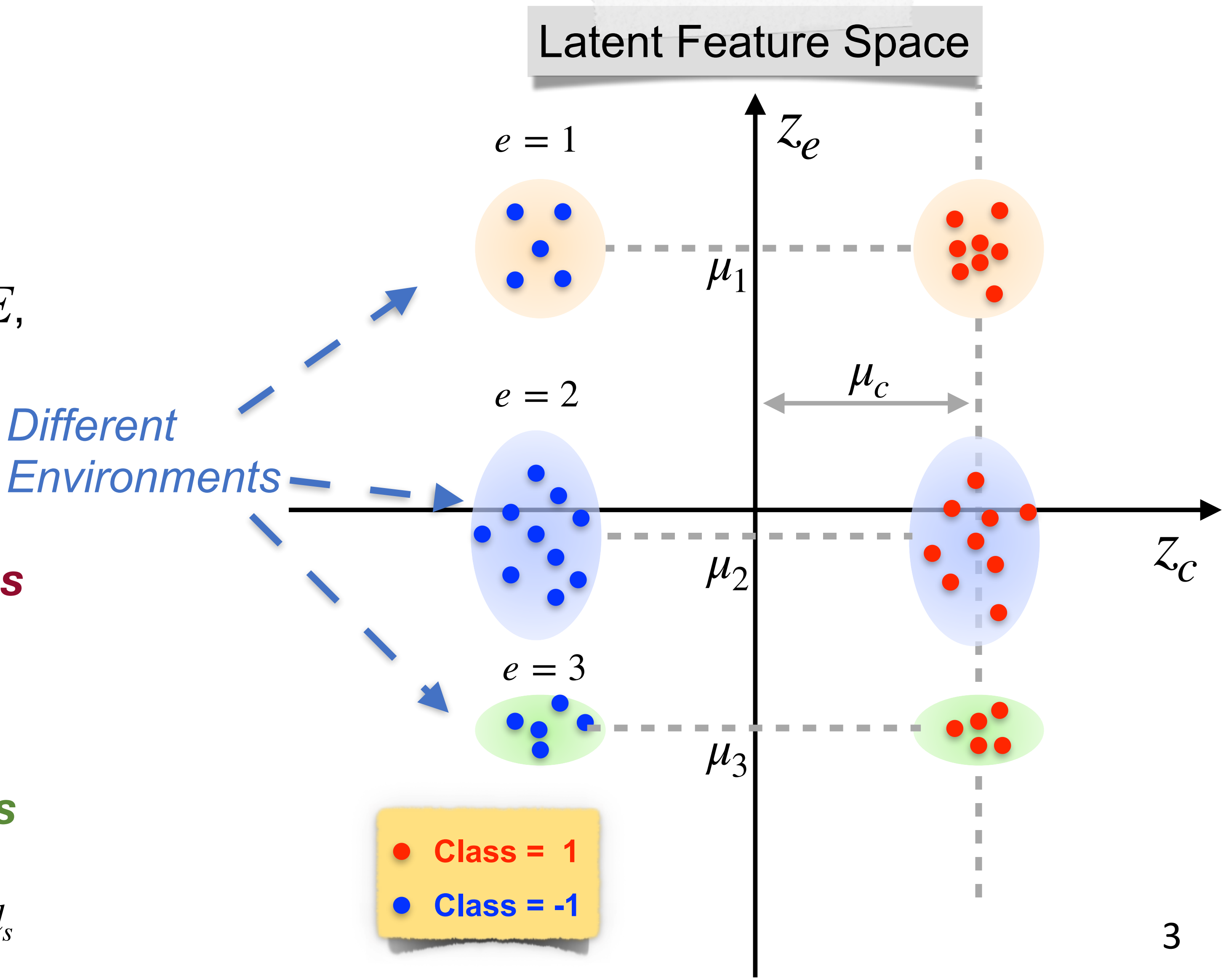
Invariant Features

$$z = \begin{bmatrix} z_c \\ z_e \end{bmatrix} \begin{cases} \boxed{z_c \sim \mathcal{N}(y\mu_c, \sigma_c^2 I) \in \mathbb{R}^{d_c}} \\ \boxed{z_e \sim \mathcal{N}(y\mu_e, \sigma_e^2 I) \in \mathbb{R}^{d_s}} \end{cases}$$

Spurious Features

$x = Az_c + Bz_e \in \mathbb{R}^d$,

where, $d = d_c + d_s$, $A = \mathbb{R}^{d \times d_c}$, $B = \mathbb{R}^{d \times d_s}$



Rosenfeld, Ravikumar and Risteski. *The Risks of Invariant Risk Minimization*. ICLR 2021.

Invariant Risk Minimization (IRM) & Optimal Invariant Predictor

IRM optimizes a bi-level objective over a feature extractor Φ and a classifier β (assumed both to be **linear** here)

$$\min_{\Phi,\beta} \sum_{e\in[E]} \mathscr{R}^e(\Phi,\beta)$$

s.t. $\beta \in \arg\min_{\beta} \mathscr{R}^e(\Phi,\beta) \quad \forall e \in [E]$

Goal of IRM → ***Optimal Invariant Predictor (OIP)***

Given $x = [A \;\; B]\begin{bmatrix} z_c \\ z_e \end{bmatrix}$, an example of OIP is:

$$\Phi^*(x) = \begin{bmatrix} z_c \\ 0 \end{bmatrix}$$ (keep invariant features only)

β^* is the optimal classifier w.r.t. $\{(\Phi^*(x), y)\}$

Assumption 1 [Mean]. For $\{\mu_e\}_{e=1}^E$, each element cannot be expressed as an affine combination of the rest.

Assumption 2 [Covariance]. There exists a pair of distinct environments $e, e' \in [E]$ s.t. $\sigma_e \neq \sigma_{e'}$.

Linear Environment Complexity: as $E > d_s$, the global optimum of IRM is guaranteed to be an optimal invariant predictor (*proved in [Rosenfeld at al. ICLR 2021]*).

Convergence Issue: IRM has no global convergence guarantee due to non-convexity.

Arjovsky, Bottou, Gulrajani, Lopez-Paz. *Invariant Risk Minimization*. arXiv 2019.

Invariant-Feature Subspace Recovery (ISR) — First Algorithm

Goal of ISR: *Find the feature subspace spanned by invariant features only.*

Algorithm 1 ISR-Mean

Input: Data of all training environments, $\{\mathcal{D}_e\}_{e\in[E]}$.

for $e = 1, 2, \ldots, E$ **do**

  Estimate the sample mean of $\{x|(x, y) \in \mathcal{D}_e, y = 1\}$ as $\bar{x}_e \in \mathbb{R}^d$

end for

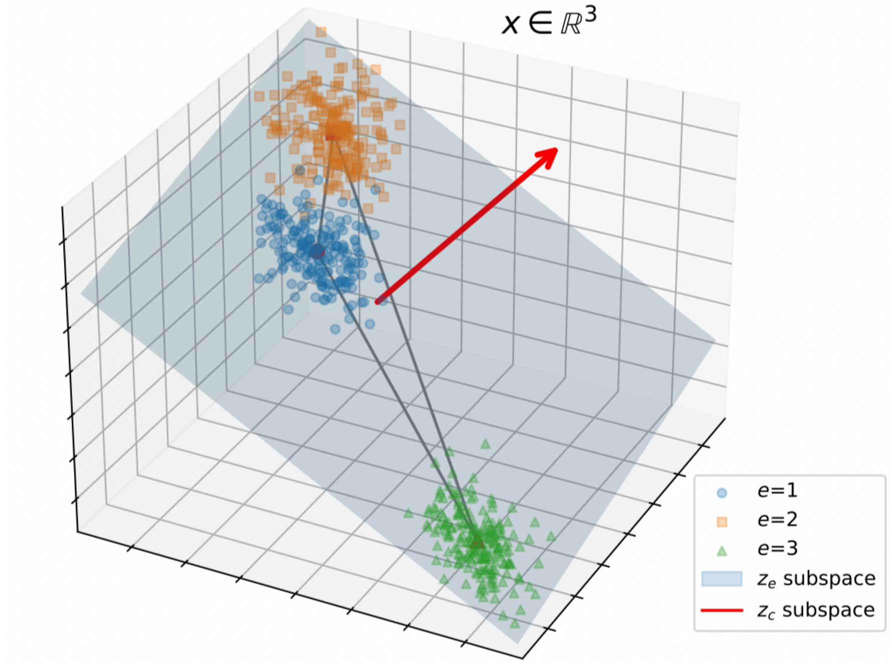
I. Construct a matrix $\mathcal{M} \in \mathbb{R}^{E\times d}$ with the e-th row as $\bar{x}_e^\top$ for $e \in [E]$

II. Apply PCA to $\mathcal{M}$ to obtain eigenvectors $\{P_1, ..., P_d\}$ with eigenvalues $\{\lambda_1, ..., \lambda_d\}$

III. Stack d_c eigenvectors with the lowest eigenvalues to obtain a transformation matrix $P' \in \mathbb{R}^{d_c\times d}$

IV. Fit a linear classifier (with $w \in \mathbb{R}^{d_c}$, $b \in \mathbb{R}$) by ERM over all training data with transformation $x \mapsto P'x$

Obtain a predictor $f(x) = w^\top P'x + b$



Linear Environment Complexity: Same as IRM.

Global Convergence: ISR-Mean enjoys global convergence guarantees.

Less Assumptions than IRM: Only Need Assumption 1 [Mean] & No need for Assumption 2 [Covariance].

Invariant-Feature Subspace Recovery (ISR) — Second Algorithm

Algorithm 2 ISR-Cov

Input: Data of all training environments, $\{\mathcal{D}_e\}_{e\in[E]}$.

for $e = 1, 2, \ldots, E$ **do**

Estimate the sample covriance of $\{x|(x, y) \in \mathcal{D}_e, y = 1\}$ as $\Sigma_e \in \mathbb{R}^{d\times d}$

end for

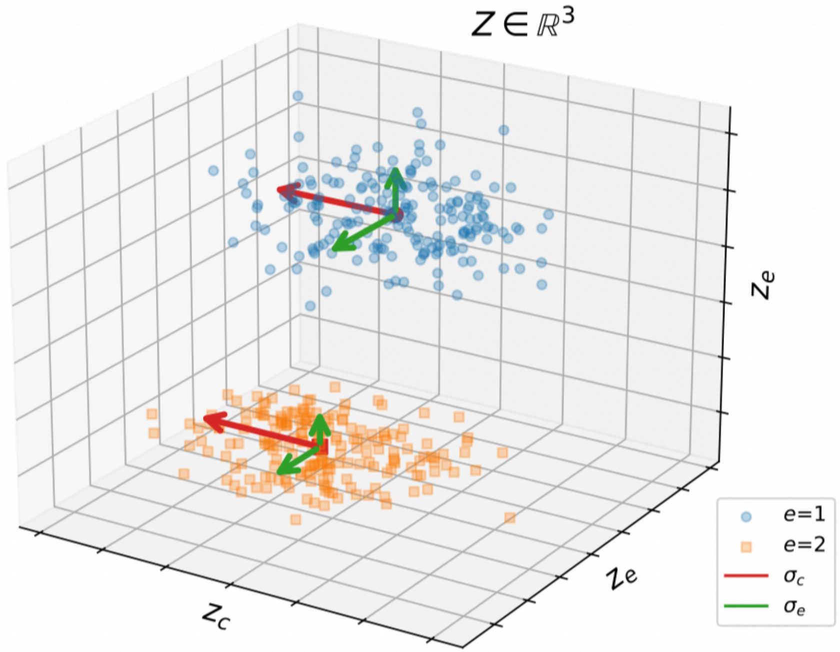
I. Select a pair of environments e_1, e_2 such that $\Sigma_1 \neq \Sigma_2$, and compute their difference, $\Delta\Sigma := \Sigma_{e_1} - \Sigma_{e_2}$

II. Eigen-decompose $\Delta\Sigma$ to obtain eigenvectors $\{P_1, ..., P_d\}$ with eigenvalues $\{\lambda_1, ..., \lambda_d\}$

III. Stack d_c eigenvectors of eigenvalues with lowest absolute values to obtain a matrix $P' \in \mathbb{R}^{d_c\times d}$

IV. Fit a linear classifier (with $w \in \mathbb{R}^{d_c}$, $b \in \mathbb{R}$) by ERM over all training data with transformation $x \mapsto P'x$

Obtain a predictor $f(x) = w^\top P'x + b$

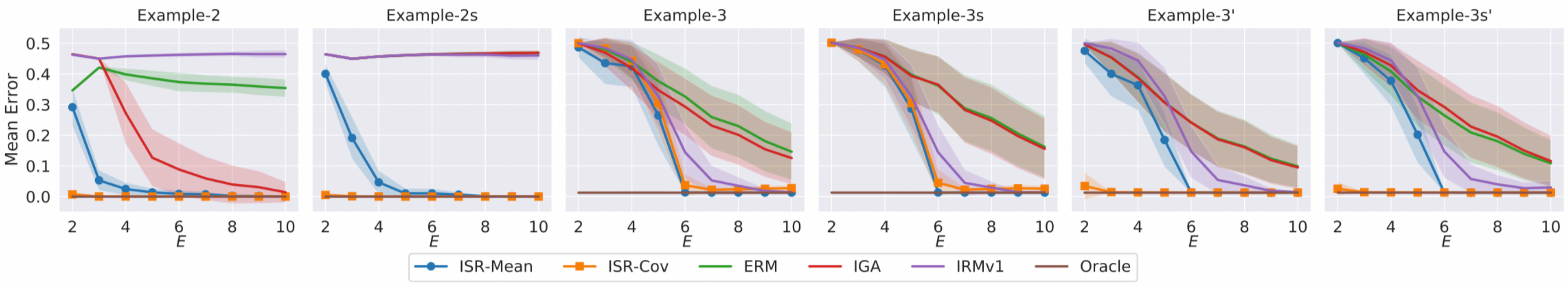


$O(1)$ **Environment Complexity:** Only needs 2 environments $\longrightarrow$ optimal invariant predictor.

Global Convergence: ISR-Cov enjoys global convergence guarantees.

Less Assumptions than IRM: Only Need Assumption 2 [Covariance] & No need for Assumption 1 [Mean].

Experiment on Synthetic Datasets: Linear Unit-Tests [Aubin et al. 2021]



Test results on Linear Unit-Tests (first 4 plots) and its variants (last 2 plots), where $d_c = 5$, $d_s = 5$, and E = 2, . . . , 10. $P(Y \mid \mu_c)$ is invariant across environments.

ISR-Mean

Linear Environment Complexity: Error is reduced to zero as $E > d_s$

Better Performance than IRM: Global convergence of ISR-Mean.

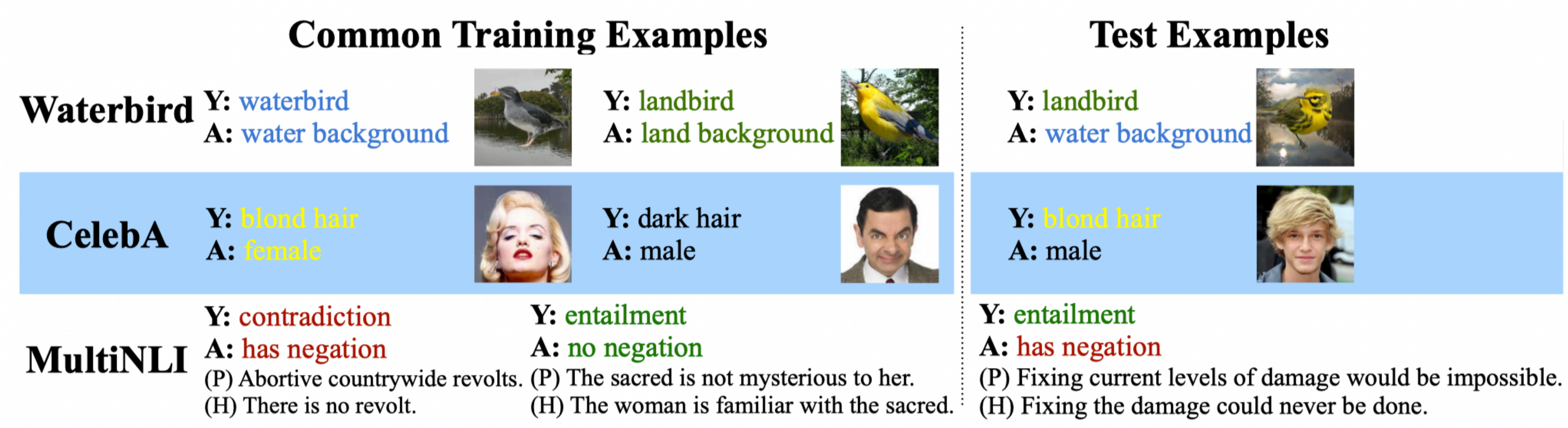
ISR-Cov

O(1) Environment Complexity: Error is reduced to zero as $E \geq 2$ for datasets that satisfy Assumption 2 [Covariance].

Aubin, Słowik, Arjovsky, Bottou, Lopez-Paz. *Linear unit-tests for invariance discovery*. 2021.

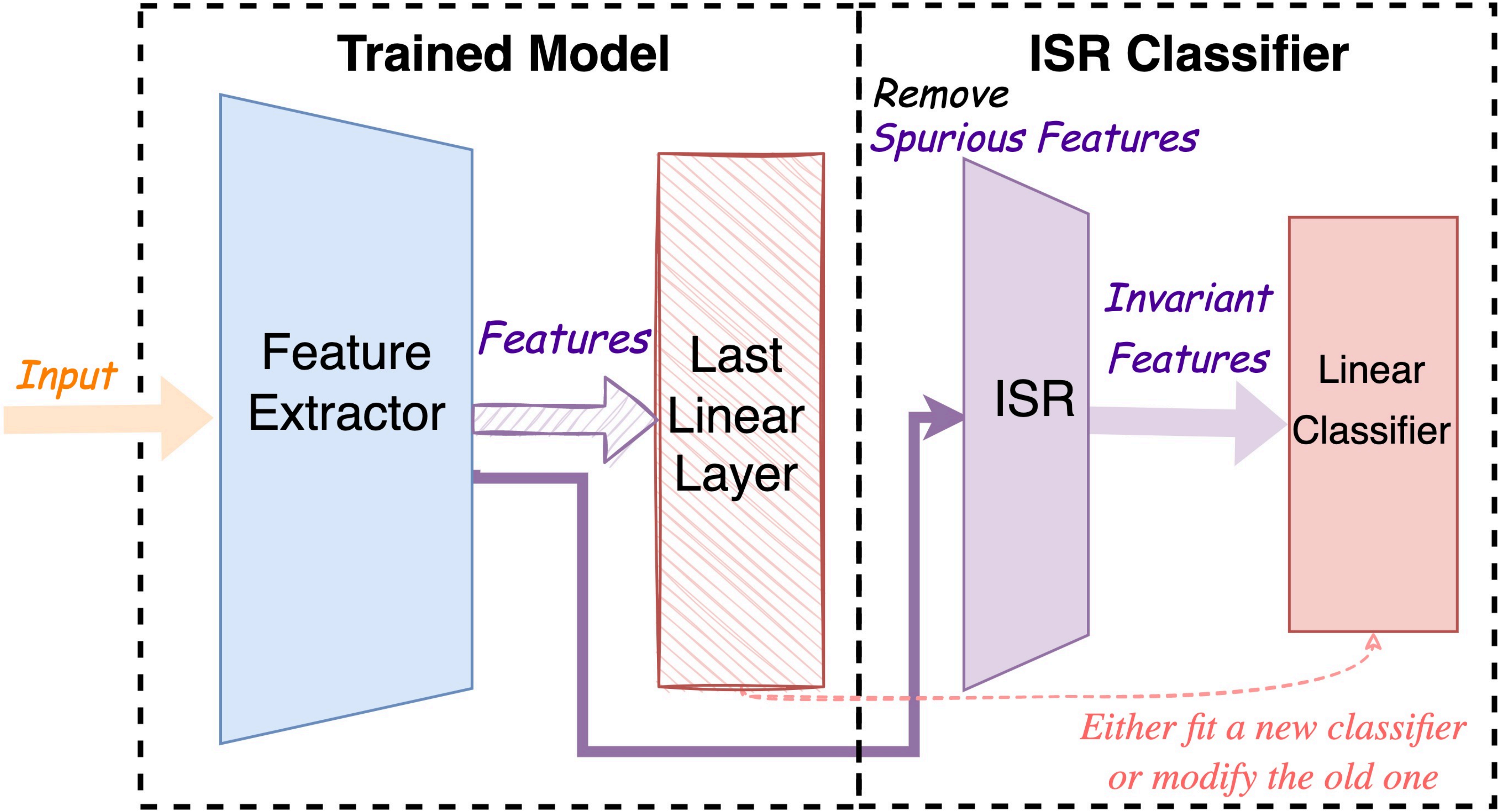
Experiments on Real Datasets

Benchmarks: three datasets used by [Sagawa et al. ICLR 2020] to study the robustness of models against spurious correlations and group shifts.



Sagawa, Koh, Hashimoto, Liang. *Distributionally Robust Neural Networks*. ICLR 2020.

ISR as Post-Processing for Trained Models: ISR Classifier on Feature Extractors



Experiments on Real Datasets

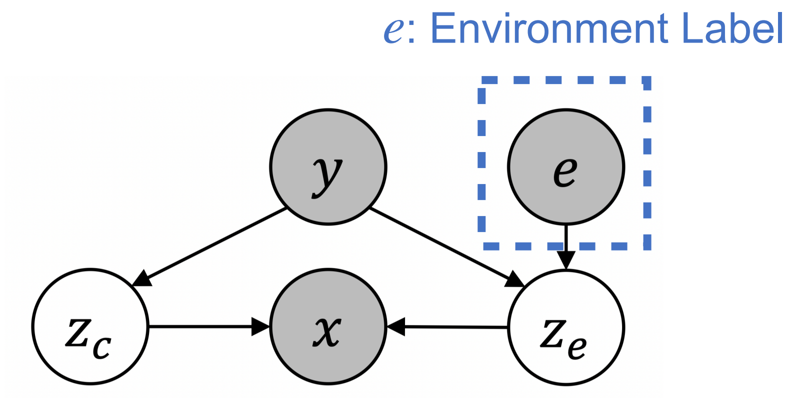
Dataset	Backbone	Algorithm	Average Accuracy			Worst-Group Accuracy		
			Original	ISR-Mean	ISR-Cov	Original	ISR-Mean	ISR-Cov
Waterbirds	ResNet-50	ERM	86.66±0.67	87.87±0.80	**90.47±0.33**	62.93±5.37	76.10±1.11	**82.46±0.55**
		Reweighting	91.49±0.46	**91.77±0.52**	91.63±0.44	87.69±0.53	88.02±0.42	**88.67±0.55**
		GroupDRO	92.01±0.33	91.74±0.35	**92.25±0.27**	90.79±0.47	90.42±0.61	**91.00±0.45**
CelebA	ResNet-50	ERM	**95.12±0.34**	94.34±0.12	90.12±2.59	46.39±2.42	55.39±6.13	**79.73±5.00**
		Reweighting	**91.45±0.50**	91.38±0.51	91.24±0.35	84.44±1.66	**90.08±0.50**	88.84±0.57
		GroupDRO	**91.82±0.27**	91.82±0.27	91.20±0.23	88.22±1.67	**90.95±0.32**	90.38±0.42
MultiNLI	BERT	ERM	**82.48±0.40**	82.11±0.18	81.28±0.52	65.95±1.65	72.60±1.09	**74.21±2.55**
		Reweighting	**80.82±0.79**	80.53±0.88	80.73±0.90	64.73±0.32	**67.87±0.21**	66.34±2.46
		GroupDRO	**81.30±0.23**	81.21±0.24	81.20±0.24	78.43±0.87	**78.95±0.95**	78.91±0.75

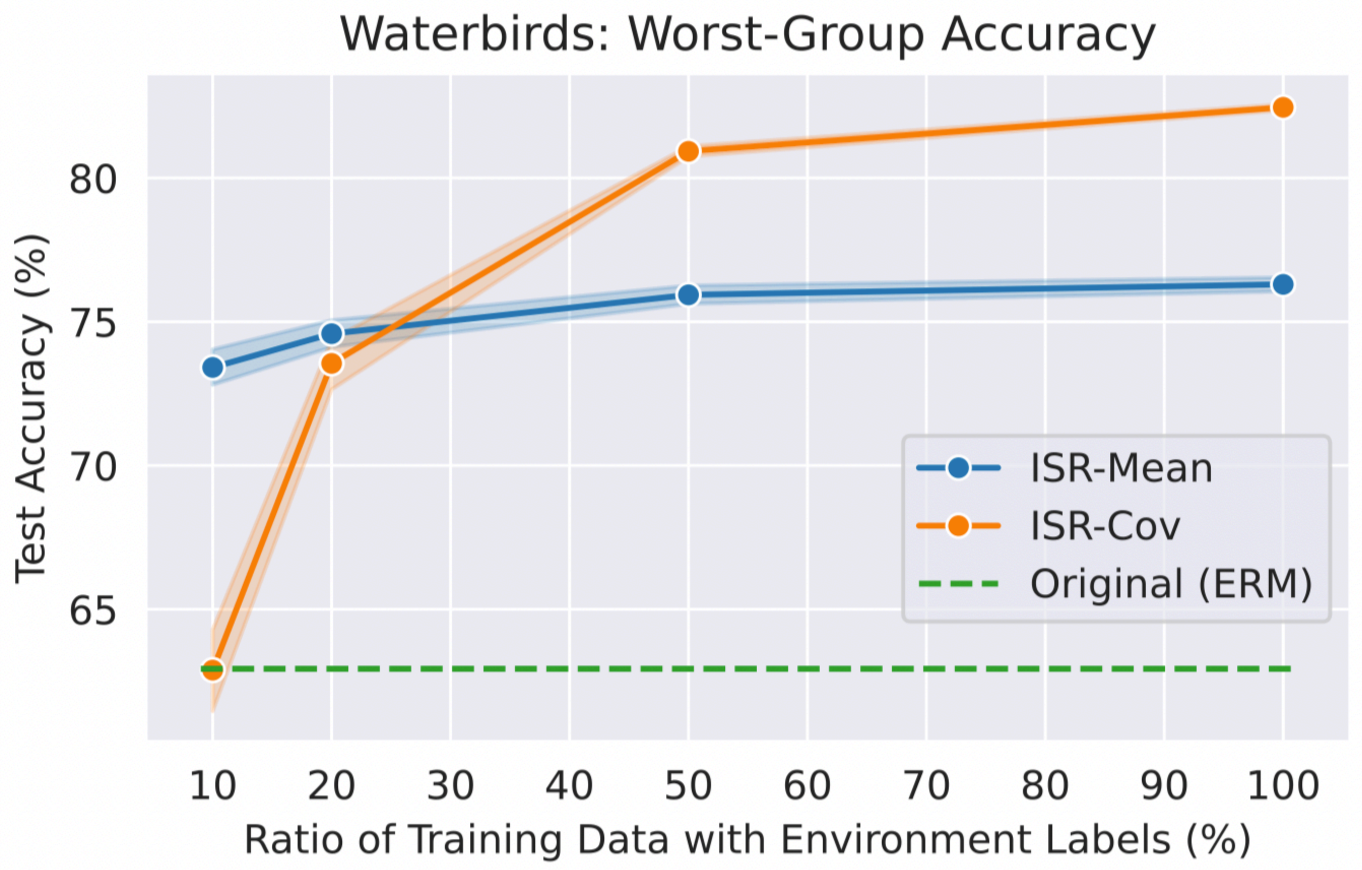
- ISR classifiers can persistently **improve the worst-group accuracy** of trained models

→ *ISR classifiers rely less on spurious features than original classifiers*

- The **average accuracy** of ISR classifiers is maintained around **the same level as** the original classifiers.

Experiments on Real Datasets: Partial Environment Labels





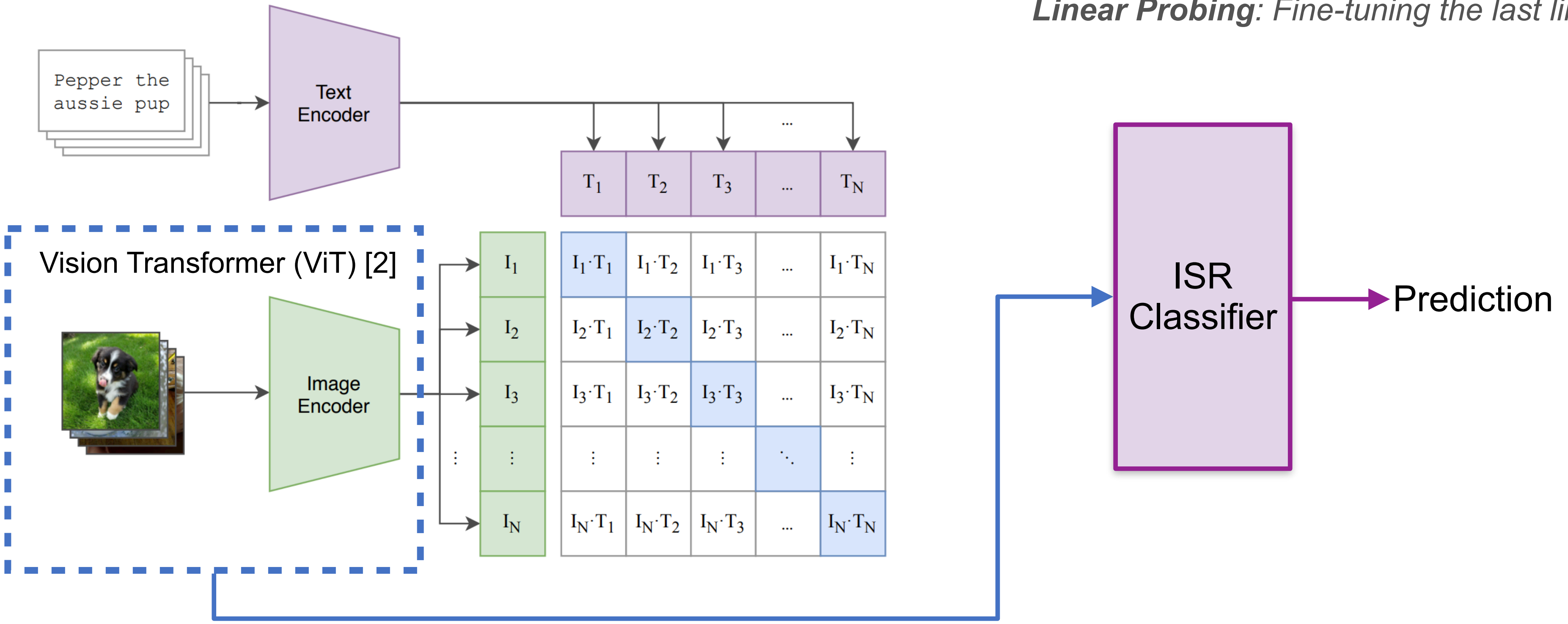
ISRs can be used in cases where only **a subset of** training samples have environment labels.

ISRs for Pre-trained Feature Extractors (No need for neural net training!)

Dataset	Backbone	Algorithm	Average Accuracy			Worst-Group Accuracy		
			Linear Probing	ISR-Mean	ISR-Cov	Linear Probing	ISR-Mean	ISR-Cov
Waterbirds	CLIP (ViT-B/32)	ERM	76.42±0.00	**90.27±0.09**	76.80±0.01	52.96±0.00	**71.75±0.39**	55.76±0.00
		Reweighting	87.38±0.09	**88.23±0.12**	88.07±0.05	82.51±0.27	**85.13±0.22**	83.33±0.00

***Linear Probing**: Fine-tuning the last linear layer only.*

Contrastive Language-Image Pre-training (CLIP) [1]



[1] Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021
[2] Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* ICLR 2021

Thank you for watching this presentation!

Code: https://github.com/Haoxiang-Wang/ISR

Contact Information:

- Haoxiang Wang: hwang264@illinois.edu
- Haozhe Si: haozhes3@illinois.edu
- Bo Li: lbo@illinois.edu
- Han Zhao: hanzhao@illinois.edu