Learning to Predict Zero:
Supervised Subspace Cancellation for Spurious Domain Robustness

Billy Peralta
Faculty of Engineering, Universidad Andrés Bello, Santiago, Chile
Accepted at the LatinX in AI Workshop at ICML 2026

Main idea: do not collapse the whole representation; cancel only a learned style-predictive subspace.

1. Motivation

Neural networks often exploit **spurious shortcuts**, such as style, color, background, or acquisition artifacts, when these factors are correlated with the task label.

A naive zero-output objective,

$$y = w^\top z \approx 0,$$

can be degenerate: the model may collapse the head or compress the embedding. Instead, we use zero prediction selectively.

Question: Can we force only the style-predictive directions to zero while preserving task-relevant information?

2. Supervised Subspace Cancellation

Given a transformed image x_i', the encoder produces

$$z_i = f_\theta(x_i') \in \mathbb{R}^{128}.$$

A domain subspace head learns

$$g_\phi(z_i) = Wz_i, \qquad W \in \mathbb{R}^{K\times 128},$$

where the rows of W are kept orthonormal:

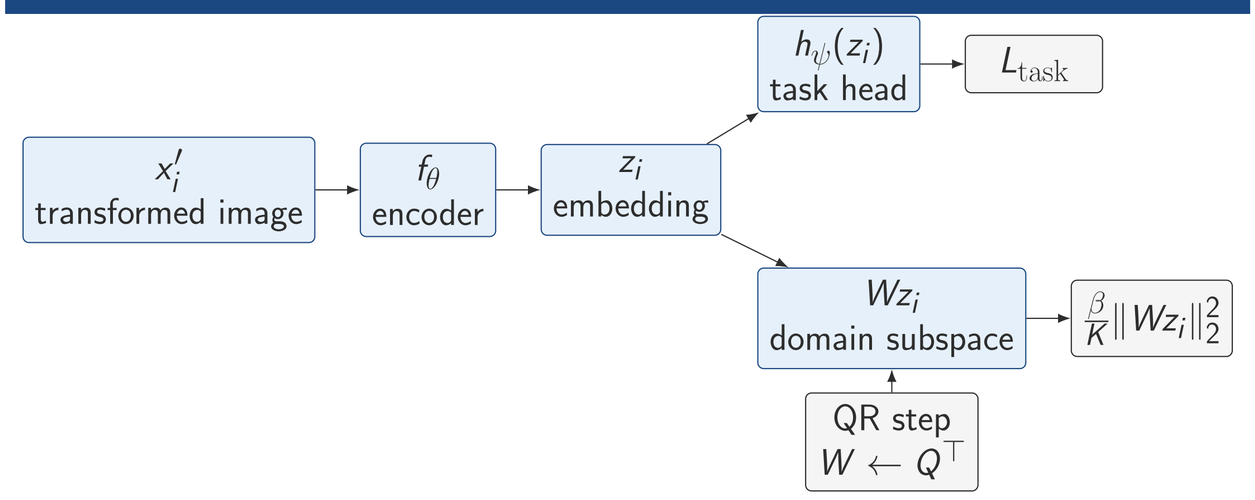
$$WW^\top = I_K.$$

The encoder is trained with the task loss plus a cancellation penalty:

$$L_{\text{ours}} = L_{\text{task}} + \frac{\beta}{K}\mathbb{E}_B\left[\|Wz_i\|_2^2\right].$$

This drives the projection onto the learned style subspace toward zero, without forcing the full representation to collapse.

3. Training Scheme



For each task mini-batch, the domain head W is updated for $S = 5$ steps on a balanced split. Then W is frozen and f_θ, h_ψ are updated using L_{ours}.

4. Controlled Spurious Setup

We evaluate on MNIST, CIFAR-10, and SVHN with synthetic binary style labels $s_i \in \{0, 1\}$.

Spurious train split: classes 0–4 are assigned one style with probability $p_{\text{corr}} = 0.99$, while classes 5–9 are assigned the opposite style.

Balanced train/test splits:

$$P(s_i = 0) = P(s_i = 1) = 0.5.$$

For MNIST:

$$T_0(x) = x, \qquad T_1(x) = \text{clip}\left(0.75(1 - x) + 0.25\epsilon, 0, 1\right).$$

The balanced test split breaks the class–style correlation used during training.

5. Linear Intuition

A cancel-only objective is not enough. In the simplified linear case, with $y = w^\top x + b$, the optimal bias centers the data:

$$b^\star = -w^\top \mu.$$

The remaining objective becomes

$$\min_w \; w^\top \Sigma w + \alpha(\|w\| - 1)^2.$$

This selects directions associated with low variance:

$$w^\star \propto u_{\min},$$

where $u_{\min}$ is the eigenvector of Σ associated with the smallest eigenvalue.

Interpretation: predicting zero can collapse toward uninformative directions unless it is paired with a task-preserving anchor. Here, Cross-Entropy preserves task information, while $\|Wz_i\|_2^2$ cancels only style-predictive directions.

6. Main Results

Lower probe accuracy means less linearly decodable style information. Since the style label is binary and balanced, **50% is chance level**.

	MNIST		CIFAR-10		SVHN	
Method	Task ↑	Probe ↓	Task ↑	Probe ↓	Task ↑	Probe ↓
Baseline	78.8	99.2	38.3	94.9	67.7	81.2
CE+SBS	74.3	99.0	36.1	95.2	64.5	82.6
Ours	78.5	**66.8**	**38.5**	**86.9**	**72.4**	**75.4**

The proposed method reduces linear probe-based style leakage while maintaining competitive task accuracy. On SVHN, task accuracy improves from 67.7% to 72.4%.

7. Sensitivity to β and K

The method shows a trade-off between invariance strength and task performance.

β	MNIST		CIFAR-10		SVHN	
	Task	Probe	Task	Probe	Task	Probe
0.5	80.6	72.7	39.8	92.6	74.0	77.0
1.0	**81.3**	61.2	39.3	90.4	**75.4**	**73.1**
2.0	80.4	**59.4**	37.5	**85.2**	74.5	74.5

Increasing β usually reduces leakage, but a very strong penalty can also remove directions useful for the main task. Similarly, too large a subspace dimension K may interfere with task-relevant features.

8. Qualitative Evidence: First-Layer Filters

The learned filters remain diverse rather than collapsed, suggesting that the cancellation objective does not simply destroy the encoder representation.

Filters are normalized per filter, clipped at $\pm 3\sigma$, and mapped to $[0, 1]$ for visualization.

9. Takeaways

- Zero prediction is useful when applied to a **selected subspace**, not to the whole representation.
- The orthonormal head W identifies style-predictive directions.
- The penalty $\|Wz_i\|_2^2$ reduces linearly decodable style leakage.
- The method does not guarantee complete removal of all style information, especially nonlinear information.
- Future work: nonlinear probes, anti-correlated tests, Waterbirds/CelebA, DANN, LEACE, and adaptive selection of K.

10. Implementation and Evaluation Details

Main setting. All main experiments use an embedding dimension $d = 128$, subspace dimension $K = 16$, cancellation weight $\beta = 2.0$, and spurious correlation $p_{\text{corr}} = 0.99$.

Optimization. The encoder and task head are trained for 5 epochs using Adam with learning rate 10^{-3} and weight decay 10^{-4}. The domain head uses Adam with learning rate 3×10^{-3}.

Alternating update. For each task mini-batch, the domain head W is updated for $S = 5$ steps on the balanced split. After each update, QR re-orthonormalization enforces $WW^\top = I_K$.

Leakage evaluation. After training, the encoder is frozen and a new linear probe is trained from scratch for 400 steps. A probe accuracy near 50% indicates chance-level prediction of the binary style label.

11. Code, References, and Acknowledgements

Code/demo: `https://zenodo.org/records/20912376`

Selected references:

- Ganin et al. Domain-adversarial training of neural networks.
- Ravfogel et al. Null It Out: guarding protected attributes by iterative nullspace projection.
- Belrose et al. LEACE: perfect linear concept erasure in closed form.
- Zbontar et al. Barlow Twins: self-supervised learning via redundancy reduction.

Acknowledgements: The author gratefully acknowledges the support of the National Center for Artificial Intelligence CENIA, Basal ANID Project FB210017, and ANID FONDECYT Regular Project No. 1241882.