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

Firstname1 Lastname1 [*1] **Firstname2 Lastname2** [*1 2] **Firstname3 Lastname3** [2] **Firstname4 Lastname4** [3]

Abstract

We study a seemingly paradoxical objective: training a neural network to map every input to the same scalar target, zero, while still learning non-trivial internal structure. A cancel-only loss enforcing $y = w^\top z \approx 0$ admits degenerate solutions (e.g., collapsing the head or compressing the embedding), so we use this principle to motivate zero-projection learning, where selected directions are forced toward zero while task-relevant structure is preserved. We extend scalar cancellation to multi-dimensional subspace cancellation by introducing an orthonormal projection head $W \in \mathbb{R}^{K \times d}$ (maintained via QR re-orthonormalization) and minimizing the projected energy $\mathbb{E}\,\|Wz\|_2^2$, which suppresses a K-dimensional subspace while preserving the orthogonal complement for task-relevant features. We instantiate this idea for supervised spurious domain cancellation with an asymmetric update schedule: a domain head is trained on a domain-balanced split to capture an injected style, after which the encoder is optimized for the main task while canceling the learned domain subspace. Across MNIST, CIFAR-10, and SVHN under controlled spurious correlations ($p_{\text{corr}} = 0.99$), our approach reduces domain leakage measured by a linear probe while maintaining competitive accuracy, and can improve generalization under shift (e.g., on SVHN). Sensitivity analyses over (K, β) further reveal a trade-off between invariance strength and main-task performance.

1. Introduction

Learning useful representations is usually associated with preserving information: an encoder is expected to retain the variations that help solve a downstream task. In this work, we study the opposite pressure. We ask whether a network can be trained to predict zero while still learning a non-trivial internal representation. At first sight, this objective appears degenerate: if every input must be mapped to the same scalar value, the model can satisfy the loss by ignoring the input, collapsing the head, or compressing the embedding.

This apparent degeneracy is precisely what makes the problem interesting. A cancel-only objective such as $y = w^\top z \approx 0$ does not by itself specify what should remain in the representation. It only specifies what should vanish. As a result, predicting zero becomes meaningful only when the zero constraint is paired with a mechanism that preserves useful structure. This leads to the central question of the paper: can zero prediction be used not as a global collapse objective, but as a controlled way to silence selected directions of a representation?

We first study this question in a simplified linear setting. The analysis shows that cancel-only learning naturally favors low-variance or null-space directions, clarifying why additional structure is needed. This motivates a nonlinear supervised instantiation in which a task loss preserves semantic information, while a separate zero-projection loss suppresses a chosen subspace. In other words, the model is not asked to make the entire representation vanish; it is asked to make a particular projection vanish.

Spurious domain learning provides a natural testbed for this idea. In many classification problems, models exploit nuisance factors such as color, texture, illumination, background, or acquisition style when these factors are correlated with the label. If such a factor is known during training, either because it is synthetically injected or available as metadata, we can learn a subspace that is predictive of it and then force the encoder to reduce its projection onto that subspace.

We instantiate this idea as supervised subspace cancellation. Given an encoder f_θ and embedding $z = f_\theta(x)$, a domain head learns an orthonormal matrix $W \in \mathbb{R}^{K \times d}$ whose rows span directions predictive of a synthetic style label. The encoder is then trained with a main-task Cross-Entropy loss together with a cancellation penalty on $\|Wz\|_2^2$. Thus, zero prediction is applied selectively: the style-predictive projec-

[*]Equal contribution [1]Department of XXX, University of YYY, Location, Country [2]Company Name, Location, Country [3]School of ZZZ, Institute of WWW, Location, Country. Correspondence to: Firstname1 Lastname1 <first1.last1@xxx.edu>.

Proceedings of the Latinx in AI workshop @ ICML, Vancouver, Canada. 2025.

tion is driven toward zero, while task-relevant information is preserved.

The paper makes three contributions. First, we provide a linear analysis of zero-output cancellation and show why unconstrained cancellation selects low-variance or degenerate directions. Second, we extend scalar cancellation to a K-dimensional orthonormal subspace learned from synthetic style labels. Third, we evaluate the resulting method on controlled spurious variants of MNIST, CIFAR-10, and SVHN, comparing against a standard Cross-Entropy baseline and a style-balanced sampling baseline.

2. Related Work

Our work connects zero-output learning with the broader problem of preserving useful structure under constrained representations. We organize the discussion around three related lines of work. First, geometry-aware architectures show that structural constraints can guide representation learning without destroying semantic information. Second, self-supervised learning demonstrates that non-standard prediction objectives can produce useful embeddings, provided that collapse is controlled. Third, collapse-avoidance methods offer principles for preventing degenerate representations. Building on these lines, we propose a supervised geometric instantiation: a domain head learns an orthonormal style-predictive subspace, and the encoder is trained to classify while driving its projection onto that subspace toward zero.

2.1. Equivariance and Geometry

Group-equivariant and steerable networks encode symmetry directly into the feature space, yielding representations that preserve global structure under rotations, reflections, and translations (Cohen & Welling, 2016; 2017; Weiler & Cesa, 2019). Recent adaptations extend these principles to Transformers, addressing shift or $E(2)$-equivariance at scale (Rojas-Gomez et al., 2024; Xu et al., 2023). Although our architecture is not explicitly equivariant, it follows a related geometric principle: useful structure is preserved by constraining how representations may vary. In our case, the constraint is not a symmetry group but an orthonormal domain subspace whose projection is explicitly suppressed.

2.2. Self-Supervised Representation Learning

Modern self-supervised learning trains encoders through predictive or alignment-based objectives without class labels. Joint-embedding predictive architectures learn by predicting masked or held-out representations in latent space (Assran et al., 2023; Bardes et al., 2024), while scaled SSL models such as DINOv2 show that high-quality visual features can emerge from large-scale pretraining (Oquab et al., 2023). Our setting is different because task labels are available, but it shares the same central concern: a non-standard prediction objective is useful only if the representation avoids collapse. We address this through a supervised task anchor, provided by Cross-Entropy, and a targeted zero-projection constraint applied only to style-predictive directions.

2.3. Output Suppression and Collapse Avoidance

Variance, covariance, and whitening penalties are standard tools for preventing representational collapse in non-contrastive learning. Barlow Twins reduces redundancy by driving cross-correlations toward identity (Zbontar et al., 2021), VICReg combines invariance with per-dimension variance and decorrelation (Bardes et al., 2022), and Whitening-MSE enforces a whitened latent space (Ermolov et al., 2021). Related methods such as BYOL avoid collapse through architectural asymmetries (Grill et al., 2020). These works motivate the principle that zero-like or alignment-based objectives must be paired with a mechanism that preserves representation structure. Rather than applying variance or whitening penalties to the whole embedding, our method targets the undesired factor directly: it learns a style-predictive subspace and suppresses only the projection onto that subspace.

Connection to this work. These strands motivate our central design principle: cancellation should be applied selectively, not globally. Our linear analysis shows why unconstrained zero-output objectives are biased toward low-variance or degenerate directions. The proposed neural instantiation turns this insight into a supervised geometric mechanism: Cross-Entropy preserves task-relevant information, while an orthonormal domain head identifies and suppresses style-predictive directions. Thus, our method translates collapse avoidance from global embedding regularization to targeted subspace cancellation.

2.4. Contributions and Novelties

Our contribution is a supervised zero-projection framework for subspace cancellation in spurious-domain robustness. First, we provide a linear analysis showing why cancel-only objectives naturally select low-variance or degenerate directions. Second, we extend scalar cancellation to a K-dimensional orthonormal domain subspace learned from synthetic style labels. Third, we introduce an asymmetric training scheme in which the domain head is trained on a balanced split and the encoder is then optimized to classify while suppressing the learned style projection. Finally, we evaluate this mechanism on controlled spurious variants of MNIST, CIFAR-10, and SVHN, including a style-balanced CE baseline.

3. Linear Regression Baseline Analysis

To better understand the collapse behavior of zero-output cancellation, we analyze the linear case where the encoder reduces to the identity mapping, i.e., $z \equiv x$. This derivation provides a closed-form solution and clarifies the implicit biases of the objective.

3.1. Formulation

We consider a scalar predictor of the form

$$y = w^\top x + b, \quad x \in \mathbb{R}^d, \tag{1}$$

with mean $\mu = \mathbb{E}[x]$ and covariance $\Sigma = \mathrm{Cov}(x)$.

We study a simplified zero-output objective with a norm constraint:

$$L(w, b) = \mathbb{E}[(w^\top x + b)^2] + \alpha(\|w\| - 1)^2. \tag{2}$$

where $\alpha > 0$ controls the strength of the norm constraint.

3.2. Optimal Bias

Expanding the quadratic term and differentiating with respect to b yields

$$\frac{\partial}{\partial b}\mathbb{E}[(w^\top x + b)^2] = 2(w^\top \mu + b) = 0, \tag{3}$$

thus

$$b^\star = -\, w^\top \mu. \tag{4}$$

In other words, the optimal bias exactly cancels the projection of w onto the mean, equivalent to centering the data.

3.3. Reduced Problem in w

Substituting $b^\star$, the expected squared output simplifies to

$$\mathbb{E}[(w^\top x + b^\star)^2] = w^\top \Sigma w. \tag{5}$$

The reduced optimization problem becomes

$$\min_{w \in \mathbb{R}^d} \; \mathcal{J}(w) = w^\top \Sigma w + \alpha(\|w\| - 1)^2. \tag{6}$$

3.4. Closed-Form Solution

Let $\Sigma = U \Lambda U^\top$ be the eigendecomposition with eigenvalues $0 \le \lambda_1 \le \cdots \le \lambda_d$ and eigenvectors $\{u_i\}$. Write $w = su$ with $\|u\| = 1$. Then

$$\mathcal{J}(s, u) = s^2(u^\top \Sigma u) + \alpha(s - 1)^2. \tag{7}$$

For fixed u, the optimal scale is

$$s^\star(u) = \frac{\alpha}{\alpha + u^\top \Sigma u}. \tag{8}$$

Minimizing over u selects the eigenvector associated with the smallest eigenvalue:

$$u^\star = u_{\min}, \quad \Sigma u_{\min} = \lambda_{\min} u_{\min}. \tag{9}$$

Thus, the closed-form optimum is

$$w^\star = \frac{\alpha}{\alpha + \lambda_{\min}}\, u_{\min}, \qquad b^\star = -(w^\star)^\top \mu. \tag{10}$$

3.5. Value at the Optimum

The expected squared output and norm are given by

$$\mathbb{E}[y^2] = (w^\star)^\top \Sigma w^\star = \frac{\alpha^2 \lambda_{\min}}{(\alpha + \lambda_{\min})^2}, \tag{11}$$

$$\|w^\star\| = \frac{\alpha}{\alpha + \lambda_{\min}}. \tag{12}$$

The minimized loss is

$$\mathcal{J}(w^\star) = \frac{\alpha \lambda_{\min}}{\alpha + \lambda_{\min}}. \tag{13}$$

3.6. Discussion

Geometric tension. The "cancel" term $w^\top \Sigma w$ pushes w toward directions of low variance, while the norm penalty enforces $\|w\| \approx 1$. The compromise results in w aligning with the eigenvector of minimum variance.

Degenerate case. If $\lambda_{\min} = 0$, there exists a non-trivial w with $\|w^\star\| = 1$ and $\mathbb{E}[y^2] = 0$. This represents a complete collapse onto a null-space direction.

Interpretation. Although predicting along eigen-min directions appears uninformative, this analysis is highly valuable: (i) it provides an analytical baseline for the loss, (ii) it highlights the risk of collapse into low-variance subspaces, and (iii) it justifies the need for an additional task-preserving or anti-collapse mechanism when moving to nonlinear encoders.

Extension to Nonlinear Networks. While the linear regression case collapses to directions of minimum variance, replacing the predictor with a multilayer perceptron (MLP) or a convolutional neural network (CNN) introduces nonlinear feature transformations that break the closed-form structure. In practice, the same cancellation objective can now interact with internal representations, potentially driving them toward degenerate states (e.g., constant activations) unless an additional task-preserving anchor or anti-collapse constraint is introduced. This motivates the supervised nonlinear instantiation studied next, where Cross-Entropy anchors task-relevant information and cancellation is restricted to a learned domain subspace.

4. Neural Instantiation for Supervised Spurious Cancellation

The linear analysis above shows that a cancel-only objective can drive the solution toward low-variance or degenerate directions. In our experiments, we instantiate the zero-prediction principle in a supervised spurious-cancellation setting rather than in a fully unsupervised representation-learning setting. Accordingly, the main-task Cross-Entropy acts as the task-preserving anchor, while the zero objective is applied only to a learned domain subspace.

4.1. Encoder and Main Task Head

Given a transformed input image x'_i, the encoder produces an embedding

$$z_i = f_\theta(x'_i) \in \mathbb{R}^{128}. \tag{14}$$

A main task head h_ψ maps this embedding to class logits,

$$\hat{y}_i = h_\psi(z_i), \tag{15}$$

and is trained with the standard Cross-Entropy loss

$$\mathcal{L}_{\text{task}} = \text{CE}(\hat{y}_i, y_i). \tag{16}$$

In practice, f_θ is a compact CNN with three convolutional blocks (Conv→BatchNorm→ReLU, with $2\times$ pooling), followed by global average pooling and a linear projection to z_i.

4.2. Domain Subspace Head

To identify directions associated with the injected style, we introduce a domain subspace head

$$g_\phi(z_i) = Wz_i, \qquad W \in \mathbb{R}^{K\times 128}, \tag{17}$$

where W has no bias. The rows of W are trained on the balanced split using the synthetic style labels s_i. With the encoder frozen, W is optimized with a multi-output binary loss,

$$\mathcal{L}_{\text{dom}} = \frac{1}{K}\sum_{k=1}^{K} \text{BCE}\big((Wz_i)_k, s_i\big), \tag{18}$$

so that its row space aligns with directions predictive of the injected style. The K outputs are treated as parallel style-predictive coordinates, not as K domain classes. After each domain-head update, we re-orthonormalize the rows of W using QR decomposition, enforcing

$$WW^\top = I_K. \tag{19}$$

Thus, W acts as an orthonormal coordinate map onto the learned domain subspace, and $W^\top W$ defines the corresponding orthogonal projector.

4.3. Subspace Cancellation Objective

Once W has captured style-predictive directions, the encoder and main task head are updated on the spurious train split by minimizing

$$\mathcal{L}_{\text{ours}} = \mathcal{L}_{\text{task}} + \frac{\beta}{K}\mathbb{E}_{\mathcal{B}}\left[\|Wz_i\|_2^2\right]. \tag{20}$$

The first term preserves task discrimination, while the second term forces the embedding to have near-zero projection onto the learned style subspace. This is the supervised form of zero prediction used in our experiments: rather than mapping the entire representation to zero, the model suppresses only the K-dimensional subspace learned as predictive of s_i.

4.4. Asymmetric Optimization

Training alternates between two steps. First, the encoder is frozen and the domain head g_ϕ is updated for S steps on the balanced loader to predict s_i, followed by QR re-orthonormalization of W. Second, W is held fixed and the encoder plus main classifier are updated using $\mathcal{L}_{\text{ours}}$. To stabilize early training, the cancellation weight β is linearly warmed up during the first epochs.

4.5. Relation to the Generic Zero-Output Objective

The generic zero-output formulation may require head-norm, variance, and decorrelation terms to prevent collapse when no supervised task signal is available. In the supervised spurious-cancellation setting studied here, these roles are handled differently: the exact orthonormal constraint on W replaces the soft head-norm penalty, and the Cross-Entropy task loss prevents global embedding collapse by requiring class-discriminative features. Therefore, our experiments use $\mathbb{E}_{\mathcal{B}}[\text{CE}(\hat{y}, y)] + \frac{\beta}{K}\mathbb{E}_{\mathcal{B}}[\|Wz\|_2^2]$.

5. Experimental Setting

5.1. Spurious Dataset Construction

For each original sample (x_i, y_i), where x_i denotes the image and y_i the class label, we generate a synthetic style/domain label $s_i \in \{0, 1\}$. The transformed image is then obtained as

$$x'_i = T_{s_i}(x_i),$$

where T_{s_i} denotes the visual transformation associated with style s_i (e.g., identity transformation for $s_i = 0$ and a combination of color inversion, contrast, brightness, or noise perturbations for $s_i = 1$). Thus, each sample used in our experiments is represented as

$$(x'_i, y_i, s_i),$$

where y_i is the main task label and s_i indicates the injected spurious style. Since the style is synthetically generated by the experimental protocol, s_i is known during training.

We construct three splits for each benchmark dataset:

- **Spurious Train Split:** The style label s_i is strongly correlated with class groups. Specifically, samples from classes 0–4 are assigned one style with probability $p_{\text{corr}} = 0.99$, whereas samples from classes 5–9 are assigned the opposite style with the same probability. This split serves as the primary training loader for the main classifier and creates a strong shortcut, since the model can partially infer the class group from the injected style.
- **Balanced Train Split:** The style label s_i is sampled independently of the class label, with $P(s_i = 0) = P(s_i = 1) = 0.5$ across all classes. This split is used to train the Domain Head g_ϕ and the final evaluation probe. Because s_i is balanced with respect to y_i, the Domain Head is encouraged to capture directions associated with the injected style rather than directions that merely reflect class identity.
- **Balanced Test Split:** Test images are also transformed using uniformly randomized styles, with $P(s_i = 0) = P(s_i = 1) = 0.5$. This provides an unbiased evaluation setting in which the spurious correlation present in the main training split no longer holds.

5.2. Evaluation Protocol: Accuracy vs. Leakage

A successful representation must classify the main task correctly without leaking domain information. We report two primary metrics on the test set:

1. **Digit Accuracy (%):** The top-1 accuracy of the main task head h_ψ.
2. **Domain Probe Accuracy (%):** To measure domain leakage, we freeze the trained encoder f_θ and train a new linear probe on the *balanced train split* for 400 steps to predict the applied style. A probe accuracy near 50% indicates optimal cancellation (random chance for a binary domain), whereas higher values indicate surviving spurious leakage in the embedding.

Code and demo. We provide an anonymized PyTorch demo implementing the synthetic style construction, asymmetric domain-head training, QR re-orthonormalization, and linear-probe evaluation: https://drive.google.com/file/d/1jljuYiNanpY54AyT6gK3CklOC9Z0lIYQ/view?usp=sharing.

6. Experiments

In this section, we present the empirical evaluation of our multi-dimensional subspace cancellation approach.We first report the main results across all benchmarks, evaluating the method's ability to reduce spurious-domain leakage across multiple random seeds. We then provide an exploratory hyperparameter sensitivity analysis for subspace dimensionality (K) and cancellation strength (β), alongside qualitative visualizations of the learned features.

6.1. Main Results: Spurious Domain Cancellation

To establish the primary efficacy of our method, we compare a standard baseline against our proposed multi-dimensional subspace cancellation model. The standard baseline corresponds to $\beta = 0$, is trained exclusively with Cross-Entropy on the spurious train split, and does not use the synthetic style labels s_i during main-task training. Based on preliminary tuning, our model is configured at $K = 16$ and $\beta = 2.0$ across all datasets. To account for variance in the spurious initialization, these core experiments are averaged over 3 random seeds. We also include a style-balanced CE baseline (CE+SBS) that uses the synthetic style labels s_i only to construct a style-balanced sampler on the spurious train split, without using a domain head, subspace cancellation, or the penalty $\beta\|Wz\|_2^2$.

As shown in Table 1, the baseline model heavily exploits the spurious shortcut, exhibiting severe domain leakage. Across all datasets, the linear probe easily identifies the spurious style in the baseline embeddings (e.g., reaching 99.2% Probe Acc on MNIST). The CE+SBS baseline shows that merely balancing the marginal frequency of styles is insufficient: the class–style shortcut remains encoded in the representation, since no explicit mechanism removes style-predictive directions.

In contrast, our proposed method consistently reduces style leakage. It achieves a substantial reduction in domain leakage, dropping the Probe Acc by over 32 percentage points on MNIST and yielding modest but consistent reductions on CIFAR-10 and SVHN. Crucially, this robust invariance does not compromise the representation quality for the primary objective. The Digit Accuracy remains highly competitive (matching the baseline in MNIST and CIFAR-10) and even demonstrates a noticeable improvement on SVHN (from 67.7% to 72.4%), suggesting that suppressing the spurious shortcut forces the encoder to learn more generalizable task features.

6.2. Hyperparameter Sensitivity (K and β)

Due to computational constraints, we perform an exploratory ablation of the subspace dimensionality ($K \in \{8, 16, 32\}$) and cancellation strength ($\beta \in \{0.5, 1.0, 2.0\}$)

Table 1. Main Results on Spurious Domain Cancellation. We compare a standard Cross-Entropy baseline, a style-balanced CE baseline (CE+SBS), and our proposed method ($K = 16, \beta = 2.0$). Results represent the mean percentage over 3 random seeds ($\pm$ standard deviation).

Method	MNIST		CIFAR-10		SVHN	
	Digit Acc ↑	Probe Acc ↓	Digit Acc ↑	Probe Acc ↓	Digit Acc ↑	Probe Acc ↓
Baseline ($\beta = 0$)	78.8 ± 2.3	99.2 ± 0.5	38.3 ± 1.5	94.9 ± 0.5	67.7 ± 0.9	81.2 ± 3.3
CE + SBS	74.3 ± 2.9	99.0 ± 0.9	36.1 ± 2.5	95.2 ± 0.6	64.5 ± 3.6	82.6 ± 1.3
Proposed (Ours)	78.5 ± 2.5	66.8 ± 8.7	38.5 ± 1.6	86.9 ± 3.0	72.4 ± 2.3	75.4 ± 3.0

using a single random seed (seed 123) across all three datasets. We decouple the analysis into two parts: varying the penalty β while keeping $K = 16$ fixed, and varying the capacity K while keeping $\beta = 1.0$ fixed.

The impact of Cancellation Strength (β). Table 2 illustrates the effect of tightening the geometric constraint. Increasing β generally reduces domain leakage, although the trend is not strictly monotonic. For instance, on MNIST, increasing β from 0.5 to 2.0 drops the Probe Acc from 72.7% to 59.4%. On CIFAR-10, it pushes the leakage down from 92.6% to 85.2%. However, this strict invariance comes with a geometric trade-off: at the highest penalty ($\beta = 2.0$), the Digit Accuracy begins to drop slightly (e.g., CIFAR-10 drops to 37.5%), as the penalty forces the embedding to squash directions that might be partially useful for the main task. The value $\beta = 1.0$ provides a reasonable intermediate trade-off, particularly evident in SVHN, where Digit Accuracy peaks at 75.4% with a low Probe Acc of 73.1%.

Table 2. Effect of Cancellation Strength (β) with fixed $K = 16$. Results are percentages (%) from a single seed.

β	MNIST		CIFAR-10		SVHN	
	Digit ↑	Probe ↓	Digit ↑	Probe ↓	Digit ↑	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

The impact of Subspace Dimensionality (K). Table 3 shows the behavior of the model when the capacity of the orthonormal projector is varied. The results demonstrate that allocating too much capacity to the domain head ($K = 32$) can be detrimental. In SVHN, expanding to $K = 32$ causes a severe performance degradation in both tasks: Digit Accuracy drops to 70.4% while Domain Leakage rebounds to 82.4% (failing to cancel the style effectively). We hypothesize that an overly large K dilutes the learning signal for the domain projection, preventing the matrix W from capturing the true low-dimensional spurious manifold. The effect of K is dataset-dependent: moderate capacities can help, but overestimating the subspace, as in $K = 32$ for SVHN, may degrade both task accuracy and leakage reduction.

Table 3. Effect of Subspace Dimensionality (K) with fixed $\beta = 1.0$. Results are percentages (%) from a single seed.

K	MNIST		CIFAR-10		SVHN	
	Digit ↑	Probe ↓	Digit ↑	Probe ↓	Digit ↑	Probe ↓
8	79.2	61.0	**40.7**	87.4	72.6	75.7
16	79.2	69.5	38.6	90.1	**75.4**	**73.1**
32	**82.6**	72.6	**40.7**	87.8	70.4	82.4

6.3. Qualitative Evidence: First-Layer Filters

To visually check for obvious low-level collapse under the cancellation penalty ($\beta = 1.0$), we inspect a subset of the first convolutional layer (`conv1`) filters. Following standard visualization practices, each filter is rendered as an image after per-filter z-score normalization, clipping at $\pm 3\sigma$, and linear mapping to the $[0, 1]$ range. For the single-channel MNIST dataset, the filters are displayed using a 'magma' colormap to highlight intensity variations, whereas for CIFAR-10 and SVHN, the 3-channel filters are rendered in standard RGB.

As shown in Figures 1, 2, and 3, the learned filters exhibit diverse and structured patterns. We observe distinct spatial frequencies, oriented edge detectors, and localized color blobs (prominent in the RGB datasets). In contrast, a truly collapsed representation would display near-constant or duplicated filters, vanishing ℓ_2 norms, or highly correlated noise. The visual diversity supports that the encoder remains non-collapsed under the cancellation objective.

6.4. Limitations

Our empirical study relies on a compact CNN and synthetic deterministic styles to isolate the mechanics of orthogonal cancellation. While this controlled setting clearly demonstrates the geometric trade-offs of our objective, extending this approach to highly complex, real-world spurious biases (e.g., Waterbirds, CelebA) and larger architectures (e.g., ResNets, ViTs) remains an open challenge. Further-

Figure 1. A subset of first-layer filters for the MNIST dataset. Single-channel weights are normalized with a z-score, clipped at $\pm 3\sigma$, and visualized using a 'magma' colormap.

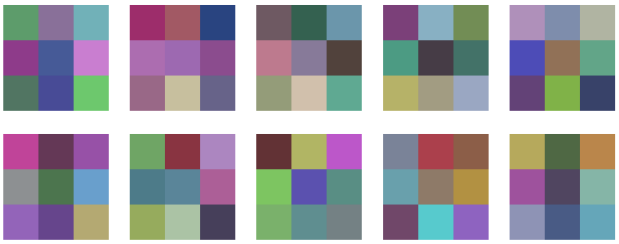

Figure 2. A subset of first-layer RGB filters for the CIFAR-10 dataset (normalized with z-score and clipped at $\pm 3\sigma$).

more, our ablation study revealed that overestimating the subspace capacity (e.g., $K = 32$) dilutes the cancellation signal, which inadvertently hurts both main-task accuracy and domain invariance. This indicates that K must currently be tuned as a discrete hyperparameter based on prior intuition about the spurious domain's complexity, highlighting the need for adaptive dimensionality allocation in future applications. The current method also assumes access to observed style labels during training; extending it to unknown spurious factors would require metadata, pseudo-domain discovery, or unsupervised subspace estimation.

7. Conclusion and Future Directions

In this work, we explored a deterministic approach to spurious domain cancellation by enforcing strict orthogonal projections during representation learning. As a non-adversarial alternative to gradient-reversal-style approaches, our method explicitly isolates style-predictive directions within a K-dimensional orthonormal subspace and penalizes its projection norm.

Our experiments provide controlled evidence that the method reduces probe-based style leakage while preserving competitive task accuracy. Furthermore, our exploratory hyperparameter analysis revealed a crucial geometric trade-off: the subspace dimensionality (K) must be sufficiently large to encapsulate the spurious manifold, yet constrained enough to prevent destructive interference with the primary semantic features (as observed when K was overparameterized to 32).

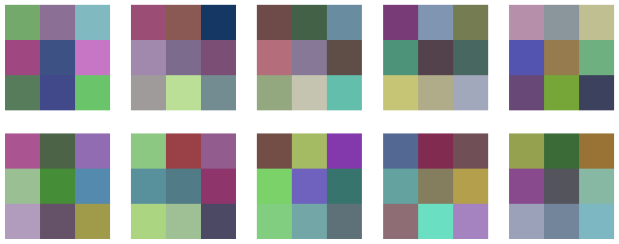

Figure 3. A subset of first-layer RGB filters for the SVHN dataset (normalized with z-score and clipped at $\pm 3\sigma$).

Future directions. Several extensions arise naturally from these findings. (i) *Dynamic K Allocation:* Investigating adaptive mechanisms, such as spectral thresholding on the domain head, to automatically discover the intrinsic dimensionality of the spurious domain during training, thereby removing the need for discrete hyperparameter sweeps. (ii) *Scaling to Complex Domains:* Testing the orthonormal cancellation mechanism on larger backbones (e.g., Vision Transformers) and real-world bias benchmarks to probe its scalability beyond synthetic deterministic styles.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., and Ballas, N. Self-supervised learning from images with a joint-embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023. URL https://arxiv.org/abs/2301.08243.

Bardes, A., Ponce, J., and LeCun, Y. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

Bardes, A., Garrido, Q., Chen, X., Rabbat, M., Ponce, J., LeCun, Y., and Ballas, N. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024. URL https://arxiv.org/abs/2404.08471.

Cohen, T. S. and Welling, M. Group equivariant convolutional networks. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *PMLR*, pp. 2990–2999, 2016. URL https://proceedings.mlr.press/v48/cohenc16.html.

Cohen, T. S. and Welling, M. Steerable cnns. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/pdf?id=rJQKYt5ll.

Ermolov, A., Siarohin, A., Sangineto, E., and Sebe, N. Whitening for self-supervised representation learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 3015–3024, 2021. URL https://proceedings.mlr.press/v139/ermolov21a.html.

Grill, J.-B., Strub, F., Altché, F., Tallec, C., Richemond, P. H., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z. D., Azar, M. G., Piot, B., Kavukcuoglu, K., Munos, R., and Valko, M. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, 2020. URL https://papers.nips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Labatut, P., Joulin, A., and Bojanowski, P. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL https://arxiv.org/abs/2304.07193.

Rojas-Gomez, R. A., Lim, T.-Y., Do, M. N., and Yeh, R. A. Making vision transformers truly shift-equivariant. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Rojas-Gomez_Making_Vision_Transformers_Truly_Shift-Equivariant_CVPR_2024_paper.html.

Weiler, M. and Cesa, G. General $e(2)$-equivariant steerable cnns. In *Advances in Neural Information Processing Systems*, 2019. URL https://papers.nips.cc/paper/2019/hash/52cb52cc2b3c5103214d5e03dba7f568-Abstract.html.

Xu, R. et al. $e(2)$-equivariant vision transformer. In *Proceedings of the 40th International Conference on Machine Learning*, volume 216 of *PMLR*, 2023. URL https://proceedings.mlr.press/v216/xu23b/xu23b.pdf.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*, pp. 12310–12320, 2021. URL https://proceedings.mlr.press/v139/zbontar21a.html.