WHEN REGULARIZATION TAKES A SHORTCUT: PREVENTING NEURAL COLLAPSE FOR OOD GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural collapse (NC) is a phenomenon observed during supervised deep training of neural networks, in which within-class representations become increasingly compact while different classes become highly separable. NC prevention methods have been shown to improve out-of-distribution (OOD) generalization. However, we observe that state-of-the-art anti-collapse regularizers act on the representation through label-agnostic projections without access to the within/between-class decomposition. Therefore, they may inadvertently trade one for the other. In this work, we investigate the interaction between within-class variability and between-class separation in the loss landscape of widely used anti-collapse regularization methods. We show theoretically that existing regularization objectives can exploit an unintended shortcut by increasing between-class separation to satisfy their optimization objectives. This limits their ability to effectively prevent NC. To address this limitation, we introduce a regularization strategy named Pinned Isotropic Entropy Regularization (PIER) that explicitly preserves within-class variability, while maintaining strong between-class separation. Our approach mitigates NC across a range of collapse regimes while preserving high in-distribution accuracy. Our extensive experiments[1] on diverse visual datasets further demonstrate improved OOD generalization across different input distributions and label spaces.

1 INTRODUCTION

Neural networks can achieve high in-distribution (ID) performance while learning representations that generalize poorly to unseen distributions (Recht et al., 2019; Taori et al., 2020; Miller et al., 2021; Djolonga et al., 2021; Fang et al., 2023; Wenzel et al., 2022; Teney et al., 2023). Moreover, models with comparable ID performance can exhibit substantially different out-of-distribution (OOD) generalization (Harun et al., 2024; Kothapalli, 2022; Masarczyk et al., 2023), suggesting that successful classification on the training distribution does not necessarily imply learning transferable representations. While some studies report a strong correlation between ImageNet accuracy and OOD accuracy (Recht et al., 2019; Taori et al., 2020; Miller et al., 2021; Djolonga et al., 2021), others (Fang et al., 2023; Wenzel et al., 2022; Teney et al., 2023; Harun et al., 2024; 2025) show that this relationship does not hold consistently across models, datasets, and training settings.

These observations have motivated a search for properties of learned representations that may explain differences in OOD generalization. One such property is NC, a characteristic phenomenon of supervised neural network training, in which learned representations and classifier weights increasingly converge toward a highly structured geometric configuration (Papyan et al., 2020). While neural collapse (NC) has been studied extensively in the context of neural network optimization and generalization, recent work has also examined its relationship with OOD generalization, revealing an association between the extent of NC and OOD generalization performance (Harun et al., 2024; Masarczyk et al., 2023). This suggests that the geometry of learned representations may provide a useful lens for understanding why models with similar ID performance can differ substantially in their ability to generalize across distributions. Recent work has linked these aspects of NC to OOD generalization, showing that models with similar ID accuracy can exhibit markedly different OOD

[1] Our code is included in the supplementary material and will be made public.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

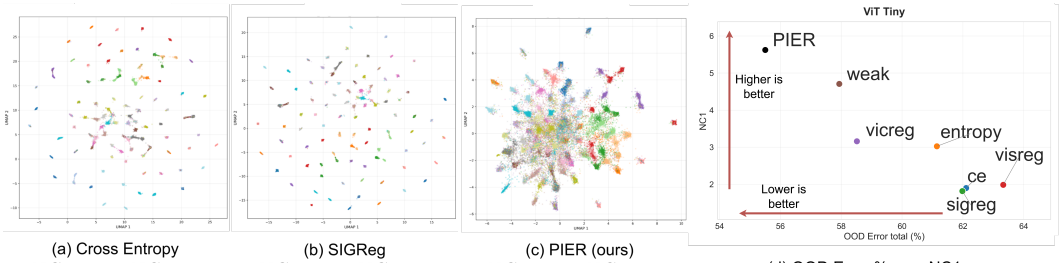

Figure 1: ViT-Tiny UMAP projection; Colors represent classes. (a–c) Comparison of cross-entropy, SIGReg (Balestriero & LeCun, 2025), and our proposed regularizer on ViT-Tiny trained on ImageNet-100. SIGReg follows a shortcut that fails to prevent the natural collapse induced by cross-entropy, whereas our method effectively mitigates within-class collapse. (d) Correlation between classification error and within-class collapse (NC1) for ViT-Tiny.

performance depending on the extent of within-class collapse during training (Harun et al., 2024; Masarczyk et al., 2023). Several elements of the training setup can alter the dynamics of collapse, e.g., image resolution, data augmentation, and number of classes. These findings suggest that the degree of representation collapse may help explain differences in OOD generalization, and have motivated regularization techniques aimed at mitigating collapse and improving OOD performance (Pan & Cao, 2023; Lin et al., 2024).

We follow the evaluation protocol of Harun et al. (2024; 2025), which considers a broad notion of OOD generalization. Specifically, these works evaluate models trained on ImageNet across eight OOD datasets that differ from ImageNet in both input distribution and label space, and study the relationship between representation collapse on ImageNet and OOD performance across these datasets. This setting provides a more direct test of whether preserving representation structure during ID training translates into improved transfer to genuinely *distinct* downstream distributions.

Under this setting, we find that existing anti-collapse regularizers do not distinguish between two fundamentally different aspects of representation geometry: within-class collapse and between-class separation. To address this limitation, we develop a theoretical framework that characterizes the interaction between within-class variability and between-class separation in supervised learning. Our analysis disentangles these two geometric dynamics in the loss landscape and reveals that existing regularizers can either overly suppress within-class variability or fail to appropriately control the trade-off between within-class dispersion and between-class angular separation, as illustrated in Figure 1. This analysis motivates a more targeted approach to representation regularization that explicitly accounts for both components of the learned representation geometry.

In this paper, we further investigate isotropic geometric regularization as a means to mitigate NC and improve OOD generalization, strengthening the connection between anti-collapse mechanisms and generalization. We identify a fundamental limitation of existing isotropic regularizers: their objectives can be satisfied when class representations are well separated in an approximately isotropic manner, even when substantial within-class collapse remains. We theoretically characterize *why* this phenomenon arises and complement our analysis with consistent empirical evidence across different settings. Our contributions can be summarized as follows.

(1) A theoretical limitation of existing regularizers. We show theoretically that existing regularization methods can satisfy their objectives by increasing between-class separation, without preventing within-class collapse.

(2) An effective anti-collapse solution. Guided by our analysis, we introduce a Pinned Isotropic Entropy Regularization (PIER) that prevents the between-class shortcut, preserves within-class variability, and mitigates NC, improving OOD generalization across different collapse regimes.

(3) Comprehensive empirical validation. We validate our theoretical analysis across diverse architectures, training settings, collapse regimes, and dataset sizes. The results corroborate our predictions in both collapsing and non-collapsing scenarios and demonstrate the effectiveness of our approach in mitigating NC and improving OOD generalization.

2 BACKGROUND

2.1 NEURAL COLLAPSE AND OOD GENERALIZATION

NC describes a geometric organization of deep neural network representations that emerges during the later stages of training. Instead of being randomly distributed, features tend to organize into a highly structured configuration where samples from the same class become increasingly concentrated, while different classes become maximally separated (Papyan et al., 2020).

In this work, we focus specifically on the two properties of NC that are most relevant to representation quality and transfer learning:

- **Feature collapse** (NC1): samples belonging to the same class progressively converge toward their class prototype, reducing within-class variance.
- **Equiangular class geometry** (NC2): class prototypes arrange into a Simplex Equiangular Tight Frame (ETF), yielding maximal between-class separation in angular space.

In open-world learning scenarios, robust transferability is particularly important, as models must adapt to previously unseen distributions. To address this challenge, a wide range of techniques have been proposed, including feature alignment (Li et al., 2018; Ahuja et al., 2021; Zhao et al., 2020; Bai et al., 2024), weights regularisation (Aminbeidokhti et al., 2026), feature disentanglement (Zhang et al., 2022), robust optimization (Rame et al., 2022; Cha et al., 2021; Krueger et al., 2021), and data augmentation (Nam et al., 2021; Aminbeidokhti et al., 2024; Zhou et al., 2020). Several key studies have investigated the factors that most strongly influence OOD generalization, identifying feature diversity as a particularly important contributor to improved robustness (Masarczyk et al., 2023; Harun et al., 2024; Kornblith et al., 2021; Fang et al., 2023; Vishniakov et al., 2023). Recent work suggests that feature compression in deeper layers, which is closely associated with NC, may hinder OOD generalization by reducing feature diversity and expressivity (Kothapalli, 2022; Masarczyk et al., 2023; Harun et al., 2024; 2025).

Prior work has investigated the relationship between NC and OOD generalization, characterizing the phenomenon and identifying several factors that influence the degree of collapse, including the number of ID classes, image resolution, data augmentation, label smoothing, and weight decay (Masarczyk et al., 2023; Harun et al., 2024; Kornblith et al., 2021; Súkeník et al., 2024; Lin et al., 2024; Rangamani & Banburski-Fahey, 2022). However, most existing studies analyze NC in pretrained models or under the same-label distribution shifts, rather than explicitly controlling NC during supervised training from scratch.

Masarczyk et al. (2023) establish a connection between NC and transferability through the tunnel effect, showing that deeper layers become increasingly specialized to the source domain, which can limit transferability. Harun et al. (2025) take a step toward directly controlling NC by regularizing deeper layers to mitigate strong collapse. Whether NC can be directly regulated during supervised training to improve generalization across shifts in both input and label distributions, however, remains largely unexplored.

Taken together, these findings motivate moving beyond characterizing NC toward directly controlling its geometry during training. Since our regularization is applied solely during ID training, without knowledge of the target OOD distributions, we seek a regularizer that preserves strong ID performance while promoting transferable representations. Following Harun et al. (2025), we therefore focus on regulating NC1 and NC2 to improve generalization under shifts in both input and label distributions.

2.2 ANTI-COLLAPSE REGULARIZATION FOR OOD GENERALIZATION

Several approaches have investigated anti-collapse regularization to improve OOD generalization (Harun et al., 2025; Lin et al., 2024; Pan & Cao, 2023; Bai et al., 2024; Wu et al., 2026; Koromilas et al., 2026). These methods broadly fall into architectural interventions and representation-level regularization. Architectural approaches, such as Pan & Cao (2023), modify normalization or network design to alter NC dynamics. In contrast, Harun et al. (2025) introduce an architecture-independent nearest-neighbor entropy regularizer that encourages diversity among neighboring representations and applies it to both CNNs and ViTs.

Other works explicitly regularize representation geometry. Hyperspherical methods improve generalization by promoting isotropy or constraining class prototypes on a hypersphere (Bai et al., 2024; Koromilas et al., 2026), but primarily consider pretrained models or shifts within the same label space. In self-supervised learning, distributional regularization toward isotropic representations has also shown promising OOD generalization under shifts in both input and label spaces (Wu et al., 2026; Balestriero & LeCun, 2025; Akbar, 2026).

Our work shares the architecture-independent, geometric regularization setup of Harun et al. (2025). Their entropy target is sound in principle, but the nearest-neighbour estimator used to optimize it is noisy: on most steps it has no same-class pair to draw on, and even when it does, it only sees one direction of within-class spread. Matching an isotropic Gaussian directly, as SIGReg (Balestriero & LeCun, 2025) does, removes this estimation noise, but the target itself is agnostic to the within/between-class decomposition. These two components can be traded off to satisfy the global isotropic Gaussian constraint. We further show that, under standard cross-entropy training dynamics, this trade-off systematically favors increasing between-class separation rather than restoring within-class variability, revealing a limitation of global distributional regularization using isotropic representation for preventing NC.

3 METHODOLOGY

Setup. We consider K-class classification on $\mathcal{X}$ with labels $\mathcal{Y} = \{1, \dots, K\}$, using a network $f_\theta : \mathcal{X} \to \mathbb{R}^K$ trained on samples from an ID distribution P_{XY}^{ID}. At an intermediate layer ℓ, let $\mathbf{z}_\ell = f_\ell(x) \in \mathbb{R}^{d_\ell}$ be the representation of x, and the marginal representation distribution is the mixture $p_\ell(\mathbf{z}) = \sum_{k=1}^K \pi_k \, p_{\ell,k}(\mathbf{z})$, with class prior $\pi_k = \Pr(Y = k)$, class means $\boldsymbol{\mu}_{\ell,k} = \mathbb{E}[\mathbf{Z}_\ell \mid Y = k]$, within-class covariances $\boldsymbol{\Sigma}_{\ell,k}$, and global mean $\boldsymbol{\mu}_\ell = \sum_k \pi_k \boldsymbol{\mu}_{\ell,k}$. By the law of total covariance, $\boldsymbol{\Sigma}_T = \boldsymbol{\Sigma}_W + \boldsymbol{\Sigma}_B$, where

$$\boldsymbol{\Sigma}_W = \sum_k \pi_k \boldsymbol{\Sigma}_{\ell,k}, \qquad \boldsymbol{\Sigma}_B = \sum_k \pi_k (\boldsymbol{\mu}_{\ell,k} - \boldsymbol{\mu}_\ell)(\boldsymbol{\mu}_{\ell,k} - \boldsymbol{\mu}_\ell)^\top, \tag{1}$$

and $\operatorname{rank}(\boldsymbol{\Sigma}_B) \le K - 1$.

Following the NC framework, *NC1* is the collapse of within-class variability, $\operatorname{Tr}(\boldsymbol{\Sigma}_{\ell,k}) \to 0$, which we measure by the scale-free ratio $\operatorname{Tr}(\boldsymbol{\Sigma}_W)/\operatorname{Tr}(\boldsymbol{\Sigma}_T)$. *NC2* is the growth of the class-mean geometry, i.e. of $\boldsymbol{\Sigma}_B$ (the separation Δ in the toy model below).

Entropy regularization: a sound target, but blind to distributional shape. Harun et al. (2025) counteract collapse by maximizing the differential entropy $H(p_\ell)$ of the layer representation. This target is not blind to NC1: since $I(\mathbf{Z}_\ell; Y) = H(\mathbf{Z}_\ell) - H(\mathbf{Z}_\ell \mid Y)$ with $H(\mathbf{Z}_\ell \mid Y) = \sum_k \pi_k H(p_{\ell,k})$,

$$H(p_\ell) = \sum_{k=1}^K \pi_k H(p_{\ell,k}) + I(\mathbf{Z}_\ell; Y), \qquad I(\mathbf{Z}_\ell; Y) \le H(\boldsymbol{\pi}) \le \log K, \tag{2}$$

and $H(p_{\ell,k}) \le \frac{1}{2} \log\det(2\pi e \, \boldsymbol{\Sigma}_{\ell,k}) \to -\infty$ as $\boldsymbol{\Sigma}_{\ell,k} \to \mathbf{0}$. Since the mutual-information term is capped by a training-independent constant, complete NC1 collapse forces $H(p_\ell) \to -\infty$: class separation cannot mask it in the population quantity, so maximizing $H(p_\ell)$ is a legitimate anti-collapse objective.

However, in practice $H(p_\ell)$ is estimated on a batch with a pooled, label-agnostic statistic. Harun et al. (2025) use the nearest-neighbour ($q = 1$) entropy estimator on the unit sphere (KoLeo),

$$\widehat{H}_{\mathrm{NN}}(\hat{\mathbf{z}}_\ell) = \frac{1}{B} \sum_{i=1}^B \log \varepsilon_1(i), \quad \varepsilon_1(i) = \min_{j \ne i} \|\tilde{\mathbf{z}}_i - \tilde{\mathbf{z}}_j\|_2, \quad \tilde{\mathbf{z}} = \frac{\mathbf{z}}{\|\mathbf{z}\|_2}, \tag{3}$$

which coincides with the Kozachenko–Leonenko estimator up to constants absorbed in $\lambda_{\mathrm{entropy}}$.

For each i, $\varepsilon_1(i)$ is a distance to a *single* neighbor, which may or may not share i's class. If it does not, the gradient carries no within-class signal. If it does, the signal is still only the scalar $\varepsilon_1(i)$, while $\boldsymbol{\Sigma}_{\ell,k}$ is a $d_\ell \times d_\ell$ PSD matrix: one pairwise distance is a rank-1 projection along whichever

direction happens to separate that pair, uninformative about the remaining $d_\ell - 1$ directions of within-class spread. What we actually want is a direct handle on that shape, not a proxy that at best sees one dimension of it. Rather than relying on a nearest-neighbor-based entropy estimator, we turn to what differential entropy maximization itself prescribes: a specific target distribution we can match directly.

The entropy-maximization target is the isotropic Gaussian.

Proposition 1 (Maximum entropy at fixed energy)**.** *Let* $\mathbf{Z}$ *have mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma}$ *in* $\mathbb{R}^d$*. Then*

$$H(\mathbf{Z}) \le \tfrac{1}{2}\log\det(2\pi e\,\boldsymbol{\Sigma}) \le \tfrac{d}{2}\log\Big(2\pi e\,\tfrac{\operatorname{Tr}\boldsymbol{\Sigma}}{d}\Big), \tag{4}$$

with equality throughout iff $\mathbf{Z} \sim \mathcal{N}\big(\boldsymbol{\mu}, \tfrac{\operatorname{Tr}\boldsymbol{\Sigma}}{d}\mathbf{I}\big)$*. Proof is in App. A.3.*

With a scale constraint, entropy is uniquely maximized by the isotropic Gaussian, whose angular marginal is uniform on the sphere (Proposition 9, App. A.3). Rather than maximizing an entropy estimate, SIGReg (Balestriero & LeCun, 2025) directly matches the empirical characteristic function of the pooled batch representations to that of $\mathcal{N}(\mathbf{0}, \mathbf{I})$ along random 1D projections. By the Cramér–Wold theorem, matching all 1D projections matches the full distribution, with an explicit target, avoiding neighbour search and its high-dimensional bias.

Global distribution matching entangles NC1 and NC2. Matching the characteristic function of the *unconditioned* batch to an isotropic Gaussian only constrains the total distribution p_ℓ. In a two-class toy model ($\boldsymbol{\mu}_{\ell,1,2} = \mp\Delta/2$ along a direction $\mathbf{v}$, common within-class variance σ^2, 1D projection along $\mathbf{v}$; for a random unit direction $\mathbf{u}$ the effective separation is $\Delta\,\mathbf{u}^\top\mathbf{v}$, which does not change the sign of the results below),

$$\mathcal{L}^{\text{glob}}_{\text{ECF}}(\Delta, \sigma^2) \propto \int w(t)\,\big|\cos(\Delta t/2)e^{-\sigma^2 t^2/2} - e^{-t^2/2}\big|^2\,dt. \tag{5}$$

Proposition 2 (NC1/NC2 entanglement, Strong-SIGReg)**.** *Let* $w > 0$ *a.e., locally bounded with at most polynomial growth (e.g.* $w(t) = e^{-t^2/2}$*), and* $\sigma^2 \in (0, 1)$*. Then* $\Delta = 0$ *is a strict local* maximum *of* $\mathcal{L}^{\text{glob}}_{\text{ECF}}$ *as a function of* Δ *(full proof in App. A.1).*

Proof sketch. $L(\Delta)$ is even in Δ. Write $A(t) = e^{-\sigma^2 t^2/2} - e^{-t^2/2}$ and $B(t) = \frac{t^2}{8}e^{-\sigma^2 t^2/2} \ge 0$. Since $\cos(\Delta t/2) = 1 - \Delta^2 t^2/8 + O(\Delta^4)$, the bracket equals $A - \Delta^2 B + O(\Delta^4)$ and

$$L(\Delta) = L(0) - 2\Delta^2 \int w(t)A(t)B(t)\,dt + O(\Delta^4). \tag{6}$$

For $\sigma^2 < 1$, $A(t) > 0$ for all $t \neq 0$, so the integral is strictly positive and the Δ^2 coefficient is negative.

For $\sigma^2 < 1$ (the training regime of interest), the loss has two descent directions: inflating σ^2 towards 1 (the intended anti-collapse effect) or growing Δ, which raises the *total* variance $\operatorname{Tr}\boldsymbol{\Sigma}_T = d\sigma^2 + \Delta^2/4$ without restoring any within-class variability, since a global matching loss only sees the total. Cross-entropy already follows the second route in the terminal phase of training (growing NC2 while shrinking NC1), so its gradient and the regularizer's align along Δ: the regularizer follows this route as a **shortcut** and is satisfied *as a by-product of cross-entropy*, and its distinctive effect (restoring σ^2) is left with a vanishing residual gradient. It worsens with K, as random 1D projections of many well-spread means are already near-Gaussian: SIGReg alone matches the cross-entropy baseline's NC1/NC2 on ImageNet-100 (Sec. 4.3).

Does the cheaper Weak-SIGReg inherit this pathology? Full-CF matching is costly, so we build on *Weak-SIGReg* (Akbar, 2026), which only matches the covariance through a random sketch $S \in \mathbb{R}^{m\times d_\ell}$ with i.i.d. $\mathcal{N}(0, 1/d_\ell)$ entries, redrawn at every training step:

$$\mathcal{L}^{\text{glob}}_{\text{Weak}}(\hat{\mathbf{Z}}) = \big\|S\,\widehat{\boldsymbol{\Sigma}}(\hat{\mathbf{Z}})\,S^\top - \mathbf{I}_m\big\|_F^2, \tag{7}$$

where $\widehat{\mathbf{\Sigma}}$ is the batch covariance. Because S is resampled, stochastic gradient descent optimizes the expected loss. For a population covariance $\mathbf{\Sigma}$, a direct computation gives

$$\mathbb{E}_S\big\|S\mathbf{\Sigma}S^\top - \mathbf{I}_m\big\|_F^2 = m\Big(\tfrac{\mathrm{Tr}\mathbf{\Sigma}}{d_\ell} - 1\Big)^2 + \tfrac{m(m+1)}{d_\ell^2}\|\mathbf{\Sigma}\|_F^2. \tag{8}$$

We evaluate equation 8 at the correct two-class covariance, the *rank-one* perturbation $\mathbf{\Sigma}_T = \sigma^2\mathbf{I}_{d_\ell} + \frac{\Delta^2}{4}\mathbf{v}\mathbf{v}^\top$; the naive isotropic model $(\sigma^2 + \Delta^2/4)\mathbf{I}$ would overestimate Δ's effect on the loss by a factor d_ℓ (App. A.2). This yields (proof in App. A.2):

Proposition 3 (Weak-SIGReg entanglement below a threshold)**.** *Consider $\mathbb{E}_S\mathcal{L}_{\mathrm{Weak}}^{\mathrm{glob}}$ evaluated at the population covariance above, and let $\alpha^\star = \dfrac{d_\ell}{d_\ell + m + 1} < 1$.*

(i) *For fixed σ^2, $\Delta = 0$ is a strict local maximum in Δ iff $\sigma^2 < \alpha^\star$, and a strict local minimum iff $\sigma^2 \geq \alpha^\star$ (degenerate at $\sigma^2 = \alpha^\star$).*

(ii) *Jointly over (σ^2, Δ), the unique global minimizer is $\Delta = 0,\ \sigma^2 = \alpha^\star$.*

Part (i) shows that the pathology of Proposition 2 reappears for Weak-SIGReg, below the threshold $\alpha^\star < 1$ ($= 0.5$ at our default sketch size; App. A.6). Part (ii) says the regularizer *in isolation* needs no separation at its minimum, at the price of a shrinkage bias ($\sigma^2 = \alpha^\star < 1$ rather than 1). Both statements concern the regularizer alone. Under cross-entropy training, which pushes Δ up and σ^2 down, the relevant equilibrium is not this fixed point and the descent direction (i) is active. Empirically, in architectures prone to collapse such as ResNets with Batch Normalization, Weak-SIGReg leaves NC1 essentially unchanged while increasing NC2 to approximately three times the cross-entropy level (Sec. 4.3).

Pinning the scale removes the trade-off. The shortcut in both propositions requires the *total* scale to be free: increasing Δ raises $\mathrm{Tr}\,\mathbf{\Sigma}_T$. If the scale is pinned, e.g. $\mathrm{Tr}\,\mathbf{\Sigma}_T = d_\ell$ (which a per-coordinate unit-variance constraint enforces), the picture changes.

Proposition 4 (Weak-SIGReg on the scale-pinned set)**.** *On $\{\mathbf{\Sigma} : \mathrm{Tr}\,\mathbf{\Sigma} = d_\ell\}$,*

$$\mathbb{E}_S\big\|S\mathbf{\Sigma}S^\top - \mathbf{I}_m\big\|_F^2 = \tfrac{m(m+1)}{d_\ell^2}\Big(d_\ell + \|\mathbf{\Sigma} - \mathbf{I}\|_F^2\Big), \tag{9}$$

which is uniquely minimized at $\mathbf{\Sigma} = \mathbf{I}$ (no shrinkage bias). In the two-class model with $\sigma^2 = 1 - \Delta^2/(4d_\ell)$, one has $\|\mathbf{\Sigma}_T - \mathbf{I}\|_F^2 = \frac{(d_\ell - 1)\Delta^4}{16\, d_\ell}$, strictly increasing in $|\Delta|$ (proof in App. A.4).

Once the scale is pinned, Weak-SIGReg becomes a pure isotropy penalty and *penalizes* class separation instead of rewarding it: it acts as a restoring force that redistributes variance from between-class to within-class directions, which is the behavior we want.

PIER Full objective. We apply the regularizer to the *unconditioned* batch. Let $\hat{\mathbf{z}}_\ell$ denote the batch of representations at layer ℓ and $\boldsymbol{\sigma}_\ell \in \mathbb{R}^{d_\ell}$ its per-coordinate standard deviation; $(1 - \boldsymbol{\sigma}_\ell)^2$ stands for $\frac{1}{d_\ell}\|\mathbf{1} - \boldsymbol{\sigma}_\ell\|_2^2$ and $\hat{\mathbf{z}}_\ell/\boldsymbol{\sigma}_\ell$ for coordinate-wise standardization. The per-layer regularizer, to be *minimized*, is

$$\mathcal{L}_\ell = \lambda_{\mathrm{scale}}(1 - \boldsymbol{\sigma}_\ell)^2 - \lambda_{\mathrm{entropy}}\, H_{\mathrm{NN}}\big(\hat{\mathbf{z}}_\ell/\boldsymbol{\sigma}_\ell\big) + \lambda_{\mathrm{sigreg}}\, \mathcal{L}_{\mathrm{Weak}}^{\mathrm{glob}}(\hat{\mathbf{z}}_\ell), \tag{10}$$

The three terms respectively control *scale*, *shape* and *isotropy*: the scale term fixes the covariance trace and removes the free-trace and shrinkage degrees of freedom; $\mathcal{L}_{\mathrm{Weak}}^{\mathrm{glob}}$ discourages variance concentration along a small number of directions; and the entropy term regularizes the angular distribution without affecting feature norms. Together, they promote a controlled-scale, isotropic representation while preventing within-class collapse. We refer to this combination of *pinned scale*, *isotropic variance*, and *entropy regularization* as **Pinned Isotropic Entropy Regularization (PIER)**. Further details and the precise implementation are given in App. A.7.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. Following Harun et al. (2025), we train on ImageNet-100 (Tian et al., 2020), and evaluate OOD generalization on eight standard benchmarks: NINCO (Bitterwolf et al., 2023), ImageNet-R (Hendrycks et al., 2021), CIFAR-100 (Krizhevsky et al., 2010), Oxford 102 Flowers (Nilsback & Zisserman, 2008), CUB-200 (Wah et al., 2011), FGVC-Aircraft (Maji et al., 2013), Oxford-IIIT Pets (Parkhi et al., 2012), and STL-10 (Coates et al., 2011).

Architectures. We consider ResNet-18, ResNet-34, ViT-Tiny, and ViT-Small, covering both CNNs and Transformers and a range of collapse regimes. We compare against Entropy (Harun et al., 2025), SIGReg (Balestriero & LeCun, 2025), VISReg (Wu et al., 2026), VICReg (Bardes et al., 2021), and Weak-SIGReg (Akbar, 2026). Our method requires no architectural modification and is therefore applicable across architectures.

Metrics. We report NC1 as the primary measure of within-class collapse and NC2 as a complementary metric to empirically assess the behavior predicted by Propositions 2 and 3. We additionally report ID accuracy and the average OOD error $\mathcal{E}_{GEN}Avg.$ across the eight benchmarks to assess classification and transfer performance.

Training. Following Harun et al. (2025), models are trained for 100 epochs with AdamW, batch size 512, cosine learning-rate decay with 5-epoch warmup, weight decay 0.05, label smoothing 0.1, and standard ImageNet augmentation. Full training details are provided in App. B.

4.2 IMPLEMENTATION DETAILS

We apply each regularizer to the pooled backbone representation $\mathbf{z}_\ell$ and minimize $\mathcal{L}_{\text{total}} = \mathcal{L}_{CE} + \beta\mathcal{R}$, where $\mathcal{R}$ is Entropy, SIGReg, Weak-SIGReg, or our proposed regularizer. We set $\beta = 0.05$ and all regularizer-specific scaling coefficients to one. We also evaluate VICReg (Bardes et al., 2021) and VISReg (Wu et al., 2026) for comparison. Further training and linear-probing details are provided in App. B.

Unless otherwise stated, we use an m-dimensional sketch with $m = d_\ell - 1$, which gives $\alpha^\star = 0.5$ for Weak-SIGReg. We also study other projection dimensions in App. D.2.

4.3 RESULTS

Results are organized as follows. We exhibit different degrees of collapse under standard cross-entropy training. We first examine our theoretical claims, starting with Proposition 2, which shows that SIGReg can take a shortcut by following the same direction as cross-entropy. We then evaluate Proposition 3 in both collapse-prone and collapse-resistant settings. Finally, we further examine our theoretical claims, as well as the efficiency and limitations of our approach.

Table 1: Comparison of ResNet18.

Model	$Acc_{\text{ID}}\uparrow$	**Neural Collapse**		$\mathcal{E}_{\text{GEN}}$ Avg. $\downarrow$
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
CE	83.12	1.10	2.47	51.57
Entropy	83.10	2.03	1.86	48.84
SIGReg	83.76	1.03	2.68	51.24
VISReg	83.00	0.77	1.53	53.40
VICReg	83.08	1.04	2.82	51.10
Weak-Sigreg	83.08	0.90	6.76	49.08
PIER	82.86	1.74	6.45	**46.93**

Table 2: Comparison of ViT-Tiny.

Model	$Acc_{\text{ID}}\uparrow$	**Neural Collapse**		$\mathcal{E}_{\text{GEN}}$ Avg. $\downarrow$
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
CE	62.86	1.90	5.01	62.11
Entropy	61.46	3.02	4.29	61.14
SIGReg	62.46	1.82	6.28	61.98
VISReg	61.18	1.98	4.09	63.33
VICReg	62.68	3.16	4.27	58.51
Weak-Sigreg	62.76	4.71	3.67	57.94
PIER	63.74	5.62	3.48	**55.50**

SIGReg induces between-class shortcut. Proposition 2 predicts that global CF matching admits a descent direction along Δ that cross-entropy already follows. In this sense, SIGReg may provide a shortcut toward a solution that is qualitatively similar to that induced by cross-entropy. This is what we observe: on both ResNet18 and ViT-Tiny, SIGReg maintains NC1 close to the cross-entropy baseline while increasing NC2. (Tables 1–2; e.g. $\mathcal{NC}_1$=1.90 vs. 1.82 and $\mathcal{NC}_2$=5.01 vs. 6.28 on ViT-Tiny), and its OOD error follows the cross-entropy baseline. This behavior is consistent across

Table 3: Comparison of ResNet34.

Model	$Acc_{\text{ID}} \uparrow$	Neural Collapse		$\mathcal{E}_{\text{GEN}}$ Avg. $\downarrow$
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
CE	84.36	0.90	2.28	54.50
Entropy	84.26	1.89	1.79	50.95
SIGReg	84.52	0.86	2.51	53.98
VISReg	84.24	0.55	1.74	57.68
VICReg	84.30	0.84	2.73	53.88
Weak-Sigreg	84.44	0.68	7.71	52.39
PIER	84.36	1.52	7.00	**48.15**

Table 4: Comparison of ViT-Small.

Model	$Acc_{\text{ID}} \uparrow$	Neural Collapse		$\mathcal{E}_{\text{GEN}}$ Avg. $\downarrow$
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
CE	62.46	1.59	6.60	60.94
Entropy	62.60	3.06	5.33	55.90
SIGReg	62.02	1.50	8.91	60.17
VISReg	60.56	2.61	4.57	57.41
VICReg	63.98	1.83	4.17	62.74
Weak-Sigreg	63.98	7.73	4.07	**52.14**
PIER	64.20	9.32	3.81	**52.38**

our experiments. Larger architectures, including ResNet-34 and ViT-Small in Tables 3–4, exhibit the same shortcut, with SIGReg and cross-entropy yielding similar NC and OOD generalization. The shortcut persists under the stronger collapse induced by SGD compared to AdamW in Tables 5–6.

Matching the covariance matrix takes a shortcut in collapsing regime. Proposition 3 shows that the pathology identified in Proposition 2 also arises for Weak-SIGReg when $\sigma^2 < \alpha^\star$ ($\alpha^\star = 0.5$ for our default sketch size; App. A.6). In this regime, Weak-SIGReg can satisfy its objective primarily by increasing between-class separation Δ, rather than counteracting within-class collapse, thereby improving NC2 without improving NC1.

Table 5: Comparison of ResNet18 trained with SGD.

Model	$Acc_{\text{ID}} \uparrow$	Neural Collapse		$\mathcal{E}_{\text{GEN}}$ Avg. $\downarrow$
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
CE	84.06	0.81	4.99	53.56
Entropy	83.78	0.98	4.84	50.97
SIGReg	84.76	0.79	5.25	53.01
VISReg	84.00	0.40	3.51	61.29
VICReg	84.20	0.81	4.99	53.32
Weak-Sigreg	84.06	0.91	6.53	50.79
PIER	83.96	1.46	6.25	**48.83**

Table 6: Comparison of ResNet34 trained with SGD.

Model	$Acc_{\text{ID}} \uparrow$	Neural Collapse		$\mathcal{E}_{\text{GEN}}$ Avg. $\downarrow$
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
CE	85.58	0.57	5.9	54.25
Entropy	85.36	0.89	5.59	52.40
SIGReg	85.7	0.58	6.35	54.18
VISReg	85.08	0.34	3.93	63.74
VICReg	85.7	0.57	5.89	54.30
Weak-Sigreg	85.24	0.85	7.33	52.11
PIER	85.88	1.33	6.87	**50.19**

Collapsing regime. We consider ResNet architectures with batch normalization and label smoothing ($\lambda = 0.1$), a setting prone to strong NC1 collapse. On ResNet-34 (Table 3), Weak-SIGReg clearly follows the shortcut: compared to Entropy, it reduces NC1 ($0.90 \rightarrow 0.68$) while increasing NC2 ($2.28 \rightarrow 7.71$), achieving worse accuracy than the Entropy baseline. Similar behavior in Table 6. ResNet-18 (Tables 1 and 5) shows the same tendency, with a larger relative gain in NC2 than NC1, although NC1 still improves slightly. In both cases, PIER yields substantially larger NC1 gains and the lowest OOD error.

No-collapse regime. We consider ViT-Tiny (Table 2), which exhibits substantially weaker collapse, as well as both architectures without label smoothing (Tables 7–8). On ResNet18 without label smoothing, Weak-SIGReg improves both NC1 and NC2 over the Entropy baseline (Table 7) together with OOD error. On ViT-Tiny without label smoothing, however, NC1 improves while NC2 *drops* sharply ($7.28 \rightarrow 3.54$); (Table 8), the reverse of the collapsing-regime pattern, with OOD error essentially unchanged. Consistent with Proposition 3, the risk of trading NC1 for NC2 is concentrated in settings that induce strong collapse. Away from that regime, Weak-SIGReg's behavior is less predictable from the toy model alone, and its outcome varies by architecture. PIER performs consistently across all settings.

Table 7: Comparison of ResNet18 without label smoothing.

Model	$Acc_{\text{ID}} \uparrow$	Neural Collapse		$\mathcal{E}_{\text{GEN}}$ Avg. $\downarrow$
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
Entropy	81.82	1.91	4.40	45.39
Weak-Sigreg	81.82	2.21	5.48	44.63
PIER	82.60	2.91	5.44	**43.73**

Table 8: Comparison of ViT-Tiny without label smoothing.

Model	$Acc_{\text{ID}} \uparrow$	Neural Collapse		$\mathcal{E}_{\text{GEN}}$ Avg. $\downarrow$
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
Entropy	61.14	4.42	7.28	56.16
Weak-Sigreg	61.62	5.88	3.54	56.08
PIER	62.76	6.06	3.58	**55.22**

PIER for Supervised Learning. Finally, PIER in Proposition 4 and Equation 10 explicitly counteracts the variance-reducing effect of CE. Even in the collapsing regime, it substantially improves NC1 over Weak-SIGReg while achieving the best OOD performance in Tables 1–3 and 5–8. Its slightly lower NC2 reflects the redistribution of variance from between-class separation toward within-class variability induced by the pinning mechanism in Proposition 4. Despite this trade-off, the resulting representations achieve the lowest OOD error across both settings.

As observed in Harun et al. (2024), larger architectures exhibit collapse behavior similar to their smaller counterparts. Consistently, ResNet-34 and ViT-Small reproduce the patterns predicted by our analysis: SIGReg largely follows the CE baseline (Proposition 2), Entropy improves disentanglement but remains insensitive to the other $d_\ell - 1$ dimensions, and Weak-SIGReg amplifies collapse whenever training pushes σ^2 below $\alpha^\star$ empirically the case for the CNN architectures with label smoothing considered here, yielding lower NC1 and higher NC2 as predicted by Proposition 3. PIER remains consistent with these observations and outperforms Entropy across the considered settings. In particular, Weak-SIGReg performs better than Entropy in non-collapsing regimes, as consistently observed in Tables 2, 7, and 4. On ViT-Small, it achieves performance comparable to PIER. At larger model scales, our method consistently outperforms entropy regularization, reducing OOD generalization error by 5.5% on ResNet-34 and by 6.3% on ViT-Small.

4.4 Ablation on the full loss.

Our ablations in Tables 9–10 show that the scale and entropy terms play complementary roles. On collapse-prone ResNet-18, scale control prevents NC1 collapse, while entropy becomes increasingly effective as representations become more class-separated. On ViT-Tiny, which naturally maintains higher within-class variability, scale control is less critical but still provides protection during early training. Overall, combining scale and entropy consistently improves performance over WeakSIGReg alone, motivating the use of both components in our final objective. More detailed ablations and analyses are provided in the App. D.

Table 9: ViT-Tiny Ablation study of the entropy and scale components.

Model	Acc_{ID} ↑	**Neural Collapse**		$\mathcal{E}_{\mathrm{GEN}}$ Avg. ↓
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
CE	62.14	1.86	4.96	62.92
Weak only	62.62	4.54	3.71	57.94
Weak + Scale	62.72	4.71	3.72	57.00
Weak + Ent	62.36	5.58	3.44	56.17
PIER	63.26	5.62	3.48	**55.50**

Table 10: ResNet-18 Ablation study of the entropy and scale components.

Model	Acc_{ID} ↑	**Neural Collapse**		$\mathcal{E}_{\mathrm{GEN}}$ Avg. ↓
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
CE	83.12	1.10	2.47	51.57
Weak	83.08	0.90	6.76	49.08
Weak + Scale	83.32	1.14	6.94	48.77
Weak + Ent	83.04	1.37	6.25	47.20
PIER	82.86	1.74	6.45	**46.90**

5 Conclusion

In this work, we introduce **Pinned Isotropic Entropy Regularization (PIER)**, which first *pins* the representation scale and then encourages isotropy and high entropy within this fixed-scale space. This prevents the regularizer from improving diversity simply by increasing between-class separation. As we show theoretically and observe empirically, existing anti-collapse regularizers can increase between-class separation without preserving within-class variability, particularly in collapsing settings. PIER addresses this issue by explicitly encouraging within-class variability while maintaining strong separation between classes. Across architectures, collapse regimes, and dataset scales, our results show that controlling both properties is important for mitigating NC and improving OOD generalization. PIER consistently reduces representation collapse and improves OOD generalization compared with existing anti-collapse regularizers, while having little to no effect on ID accuracy. At ImageNet-1K scale, however, the improvements are more modest, suggesting that anti-collapse regularization alone has limited potential. We find that PIER is most effective when combined with complementary techniques, such as data augmentation, rather than used as a replacement for them. Our approach is architecture-agnostic and can potentially extend to tasks such as object detection and regression. An interesting direction for future work is to investigate whether the more diverse representations encouraged by PIER can improve forward transfer in continual learning.

AI USE STATEMENT

In this work, we used generative AI tools to assist with writing, grammar correction, and the formulation of mathematical proofs, as well as to help present our ideas in a clear and accessible manner for the reader. We did not use generative AI tools to conduct the literature review, run experiments, develop the research methodology, or generate the research findings.

All AI-assisted content was carefully reviewed by the authors. The mathematical findings and core analyses were developed by the authors and subsequently refined with the assistance of LLMs. In some cases, the use of LLMs helped us further develop and extend our analysis beyond our initial expectations.

Any mathematical content generated or suggested by an LLM was independently verified by two authors and corrected by the authors when necessary. We take full responsibility for the final content of this work, including all text, claims, mathematical results, and other artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

In this work, we ensure full reproducibility across random seeds. In particular, all regularization methods involving randomly sketched projections are fully reproducible when using the same seed. Our code is included in the supplementary material and will also be made publicly available, along with the model weights used throughout our study, to facilitate further comparison and benchmarking. All neural network architectures are implemented using the 'timm' framework to further ensure reproducibility of the deep neural network experiments.

REFERENCES

Kartik Ahuja, Ethan Caballero, Dinghuai Zhang, Jean-Christophe Gagnon-Audet, Yoshua Bengio, Ioannis Mitliagkas, and Irina Rish. Invariance principle meets information bottleneck for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 34:3438–3450, 2021.

Habibullah Akbar. Weak-sigreg: Covariance regularization for stable deep learning. *arXiv preprint arXiv:2603.05924*, 2026.

Masih Aminbeidokhti, Fidel A Guerrero Pena, Heitor Rapela Medeiros, Thomas Dubail, Eric Granger, and Marco Pedersoli. Domain generalization by rejecting extreme augmentations. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2204–2214. IEEE, 2024.

Masih Aminbeidokhti, Heitor Rapela Medeiros, Srikanth Muralidharan, Eric Granger, and Marco Pedersoli. High-rate mixout: Revisiting mixout for robust domain generalization. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 3803–3812. IEEE, 2026.

Haoyue Bai, Yifei Ming, Julian Katz-Samuels, and Yixuan Li. Hypo: Hyperspherical out-of-distribution generalization. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 49188–49214, 2024.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. *arXiv preprint arXiv:2105.04906*, 2021.

Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. *arXiv preprint arXiv:2306.00826*, 2023.

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. Swad: Domain generalization by seeking flat minima. *Advances in Neural Information Processing Systems*, 34:22405–22418, 2021.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Josip Djolonga, Jessica Yung, Michael Tschannen, Rob Romijnders, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Matthias Minderer, Alexander D'Amour, Dan Moldovan, et al. On robustness and transferability of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16458–16468, 2021.

Alex Fang, Simon Kornblith, and Ludwig Schmidt. Does progress on imagenet transfer to real-world datasets? *Advances in Neural Information Processing Systems*, 36:25050–25080, 2023.

Md Yousuf Harun, Jhair Gallardo, and Christopher Kanan. Controlling neural collapse enhances out-of-distribution detection and transfer learning. In *International Conference on Machine Learning*, pp. 22149–22176. PMLR, 2025.

Yousuf Harun, Kyungbok Lee, Jhair Gallardo, Giri Krishnan, and Christopher Kanan. What variables affect out-of-distribution generalization in pretrained models? *Advances in Neural Information Processing Systems*, 37:56479–56525, 2024.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329. IEEE, 2021.

Simon Kornblith, Ting Chen, Honglak Lee, and Mohammad Norouzi. Why do better loss functions lead to less transferable features? *Advances in Neural Information Processing Systems*, 34: 28648–28662, 2021.

Panagiotis Koromilas, Theodoros Giannakopoulos, Mihalis A. Nicolaou, and Yannis Panagakis. Neural collapse by design: Learning class prototypes on the hypersphere, 2026.

Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *arXiv preprint arXiv:2206.04041*, 2022.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 5(4):1, 2010.

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pp. 5815–5826. PMLR, 2021.

Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5400–5409. Ieee, 2018.

Yong Lin, Lu Tan, Yifan Hao, Honam Wong, Hanze Dong, Weizhong Zhang, Yujiu Yang, and Tong Zhang. Spurious feature diversification improves out-of-distribution generalization. In *International Conference on Learning Representations*, volume 2024, pp. 11510–11566, 2024.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Wojciech Masarczyk, Mateusz Ostaszewski, Ehsan Imani, Razvan Pascanu, Piotr Miłoś, and Tomasz Trzcinski. The tunnel effect: Building data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 36:76772–76805, 2023.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pp. 7721–7735. PMLR, 2021.

Hyeonseob Nam, HyunJae Lee, Jongchan Park, Wonjun Yoon, and Donggeun Yoo. Reducing domain gap by reducing style bias. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 8686–8695. IEEE, 2021.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Leyan Pan and Xinyuan Cao. Towards understanding neural collapse: The effects of batch normalization and weight decay. *arXiv preprint arXiv:2309.04644*, 2023.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? *Advances in neural information processing systems*, 34:12116–12128, 2021.

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International conference on machine learning*, pp. 18347–18377. PMLR, 2022.

Akshay Rangamani and Andrzej Banburski-Fahey. Neural collapse in deep homogeneous classifiers and the role of weight decay. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4243–4247. IEEE, 2022.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Amir M Sarfi, Zahra Karimpour, Muawiz Chaudhary, Nasir M Khalid, Mirco Ravanelli, Sudhir Mudur, and Eugene Belilovsky. Simulated annealing in early layers leads to better generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 20205–20214, 2023.

Peter Súkeník, Christoph Lampert, and Marco Mondelli. Neural collapse vs. low-rank bias: Is deep neural collapse really optimal? *Advances in Neural Information Processing Systems*, 37: 138250–138288, 2024.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. Id and ood performance are sometimes inversely correlated on real-world datasets. *Advances in Neural Information Processing Systems*, 36:71703–71722, 2023.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020.

Kirill Vishniakov, Zhiqiang Shen, and Zhuang Liu. Convnet vs transformer, supervised vs clip: Beyond imagenet accuracy. *arXiv preprint arXiv:2311.09215*, 2023.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. *Advances in Neural Information Processing Systems*, 35:7181–7198, 2022.

Haiyu Wu, Randall Balestriero, and Morgan Levine. Visreg: Variance-invariance-sketching regularization for jepa training. *arXiv preprint arXiv:2606.02572*, 2026.

Hanlin Zhang, Yi-Fan Zhang, Weiyang Liu, Adrian Weller, Bernhard Schölkopf, and Eric P Xing. Towards principled disentanglement for domain generalization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8014–8024. IEEE, 2022.

Shanshan Zhao, Mingming Gong, Tongliang Liu, Huan Fu, and Dacheng Tao. Domain generalization via entropy regularization. *Advances in neural information processing systems*, 33:16096–16107, 2020.

Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. *Advances in Neural Information Processing Systems*, 35:31697–31710, 2022.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pp. 561–578. Springer, 2020.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34:29820–29834, 2021.

A Full Derivations: NC1/NC2 Entanglement for Strong- and Weak-SIGReg

This appendix gives the complete derivations summarized in Sec. 3: the entanglement result for Strong-SIGReg (App. A.1), the exact expected-loss formula and the corrected two-class analysis for Weak-SIGReg (App. A.2), and an extended discussion (App. A.6).

A.1 Strong-SIGReg: global CF matching entangles NC1 and NC2

Let a be a random unit projection direction and, for a sample x_i of class k, write $u_i = a^\top \mathbf{z}_{\ell,i}$. Under the Gaussian toy model $u_i \sim \mathcal{N}(\mu_k, \sigma_k^2)$ with $\mu_k := a^\top \boldsymbol{\mu}_{\ell,k}$, $\sigma_k^2 := a^\top \boldsymbol{\Sigma}_{\ell,k} a$, the characteristic function of a single class is $\varphi_k(t) = e^{i\mu_k t} e^{-\sigma_k^2 t^2/2}$, so $|\varphi_k(t)| \to 1$ as $\sigma_k^2 \to 0$: a collapsed class is characteristically indistinguishable from a Dirac mass, and any loss penalizing deviation from $\varphi_G(t) = e^{-t^2/2}$ resists collapse by construction:

$$\mathcal{L}_{\mathrm{ECF}}^{\mathrm{glob}} \propto \int w(t) \left| \widehat{\varphi}_{\mathrm{mix}}(t) - \varphi_G(t) \right|^2 dt. \tag{11}$$

Consider two equiprobable classes with common variance σ^2 and symmetric means $\mu_1 = -\Delta/2,\ \mu_2 = \Delta/2$; the mixture CF is $\varphi_{\mathrm{mix}}(t) = \cos(\Delta t/2) e^{-\sigma^2 t^2/2}$.

Proposition 5 (Resistance to collapse)**.** *With* $w(t) = e^{-t^2/2}$,

$$\left. \frac{\partial \mathcal{L}_{\mathrm{ECF}}^{\mathrm{glob}}}{\partial \sigma^2} \right|_{\sigma^2 = 0} \propto \int t^2 \cos\Big(\frac{\Delta t}{2}\Big) \Big[e^{-t^2/2} - \cos\Big(\frac{\Delta t}{2}\Big) \Big] e^{-t^2/2}\, dt$$

is strictly negative in both the $\Delta \to 0$ *and* $\Delta \to \infty$ *limits.*

Proof. $\Delta \to 0$**.** The integrand converges to $t^2(e^{-t^2} - e^{-t^2/2})$, strictly negative for $t \neq 0$ since $e^{-t^2} < e^{-t^2/2}$. $\Delta \to \infty$**.** Expanding the integrand as $t^2 \cos(\Delta t/2) e^{-t^2} - t^2 \cos^2(\Delta t/2) e^{-t^2/2}$, the first term vanishes by Riemann–Lebesgue against $t^2 e^{-t^2}$. Writing $\cos^2(\Delta t/2) = \frac{1}{2}(1 + \cos(\Delta t))$,

the second term splits into $\frac{1}{2}\int t^2 e^{-t^2/2}dt$ plus an oscillatory term vanishing by Riemann–Lebesgue, giving the limit $-\frac{1}{2}\int t^2 e^{-t^2/2}dt < 0$. □

So $\mathcal{L}_{\mathrm{ECF}}^{\mathrm{glob}}$ resists $\sigma^2 \to 0$ regardless of Δ – but it can equally be minimized by moving along the Δ axis instead, which is where NC2 entanglement enters.

Remark. The limits $\Delta \to 0, \infty$ in Proposition 5 are not a coincidence of the endpoints: numerical quadrature of the same integral for $w(t) = e^{-t^2/2}$ gives -0.52, -0.40 and -1.65 at $\Delta = 1, 2, 5$ respectively (against exactly $\sqrt{\pi}/2 - \sqrt{2\pi} \approx -1.62$ at $\Delta = 0$ and $-\frac{1}{2}\sqrt{2\pi} \approx -1.25$ as $\Delta \to \infty$), so the anti-collapse gradient remains negative across the whole separation range tested.

Proposition 6 (NC1/NC2 entanglement)**.** *Let $w(t) > 0$ a.e., locally bounded with at most polynomial growth (so that all integrals below converge, e.g. $w(t) = e^{-t^2/2}$), and let $\sigma^2 \in (0, 1)$. Then $\Delta = 0$ is a strict local maximum of $\mathcal{L}_{\mathrm{ECF}}^{\mathrm{glob}}(\Delta, \sigma^2)$ as a function of Δ.*

Proof. $L(\Delta)$ is even in Δ, so $\partial_\Delta L|_{\Delta=0} = 0$. Expanding $\cos(\Delta t/2) = 1 - \Delta^2 t^2/8 + O(\Delta^4)$ and writing $A(t) := e^{-\sigma^2 t^2/2} - e^{-t^2/2}$, $B(t) := \frac{t^2}{8} e^{-\sigma^2 t^2/2} \geq 0$, we get $[\cos(\Delta t/2) e^{-\sigma^2 t^2/2} - e^{-t^2/2}]^2 = A(t)^2 - 2\Delta^2 A(t)B(t) + O(\Delta^4)$, hence

$$L(\Delta) = L(0) - 2\Delta^2 \int w(t)A(t)B(t)\,dt + O(\Delta^4). \tag{12}$$

For $\sigma^2 < 1$, $A(t) > 0\ \forall t \neq 0$, $B(t) \geq 0$, $w(t) > 0$, so the integral is strictly positive and $\Delta = 0$ is a strict local maximum. The $O(\Delta^4)$ remainder is uniform in t after integration: since $|\cos x - 1 + x^2/2| \leq x^4/24$ and w has at most polynomial growth while A and B decay exponentially, dominated convergence applies, so a strictly negative Δ^2 coefficient makes $\Delta = 0$ a strict local maximum. □

A.2 WEAK-SIGREG: EXACT EXPECTED-LOSS FORMULA AND THE CORRECT TOY MODEL

Akbar (2026) replace full-CF matching by a covariance-only constraint, via a random sketch $S \in \mathbb{R}^{m\times d}$ with i.i.d. $\mathcal{N}(0, 1/d)$ entries ($d := d_\ell$):

$$\mathcal{L}_{\mathrm{Weak}}(\mathbf{Z}) = \left\| S\,\widehat{\mathbf{\Sigma}}(\mathbf{Z})\,S^\top - \mathbf{I}_m \right\|_F^2, \qquad \widehat{\mathbf{\Sigma}}(\mathbf{Z}) = \tfrac{1}{B-1}\mathbf{Z}^\top\mathbf{Z} \tag{13}$$

Lemma 1 (Expected sketched loss)**.** *For any symmetric $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$,*

$$\mathbb{E}_S \left\| S\mathbf{\Sigma}S^\top - \mathbf{I}_m \right\|_F^2 = \frac{m(m+1)}{d^2}\,\mathrm{tr}(\mathbf{\Sigma}^2) + \frac{m}{d^2}\big(\mathrm{tr}\,\mathbf{\Sigma}\big)^2 - \frac{2m}{d}\,\mathrm{tr}\,\mathbf{\Sigma} + m. \tag{14}$$

Proof. Write row s_j of S as $s_j = d^{-1/2} w_j$, $w_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ i.i.d. For $i \neq j$: independence and $\mathbb{E}[w_j w_j^\top] = \mathbf{I}_d$ give

$$\mathbb{E}[(s_i^\top \mathbf{\Sigma} s_j)^2] = \mathbb{E}_{s_i}\left[s_i^\top \mathbf{\Sigma}\, \mathbb{E}[s_j s_j^\top]\, \mathbf{\Sigma} s_i \right] = \tfrac{1}{d}\,\mathbb{E}[s_i^\top \mathbf{\Sigma}^2 s_i] = \tfrac{1}{d^2}\mathrm{tr}(\mathbf{\Sigma}^2),$$

summed over the $m(m-1)$ off-diagonal entries: $m(m-1)\,\mathrm{tr}(\mathbf{\Sigma}^2)/d^2$. For $i = j$: $s_i^\top \mathbf{\Sigma} s_i = d^{-1} w_i^\top \mathbf{\Sigma} w_i$, and using $\mathbb{E}[(w^\top A w)^2] = (\mathrm{tr}\,A)^2 + 2\,\mathrm{tr}(A^2)$ for $w \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$,

$$\mathbb{E}[(s_i^\top \mathbf{\Sigma} s_i - 1)^2] = \tfrac{1}{d^2}\left[(\mathrm{tr}\,\mathbf{\Sigma})^2 + 2\,\mathrm{tr}(\mathbf{\Sigma}^2)\right] - \tfrac{2}{d}\,\mathrm{tr}\,\mathbf{\Sigma} + 1,$$

summed over m diagonal entries. Adding the off-diagonal contribution and collecting terms ($m(m-1) + 2m = m(m+1)$) gives equation 14. □

Equation equation 14 depends on $\mathbf{\Sigma}$ *only* through $\mathrm{tr}(\mathbf{\Sigma})$ and $\mathrm{tr}(\mathbf{\Sigma}^2)$ – a consequence of the rotational invariance of S. By Cauchy–Schwarz, $\mathrm{tr}(\mathbf{\Sigma}^2) \geq (\mathrm{tr}\,\mathbf{\Sigma})^2/d$ with equality iff $\mathbf{\Sigma}$ is isotropic, so equation 14 is uniquely minimized over all PSD $\mathbf{\Sigma}$ at $\mathbf{\Sigma}^\star = \alpha^\star \mathbf{I}_d$,

$$\alpha^\star = \frac{d}{d+m+1} \in (0, 1). \tag{15}$$

Why the naive isotropic toy model is wrong. Directly transplanting Sec. A.1's toy model would set $\mathbf{\Sigma}_T = (\sigma^2 + \Delta^2/4)\mathbf{I}_d$, i.e. treat the between-class separation as inflating *every* one of the d eigenvalues equally. This is inconsistent with the two-class geometry it claims to describe: with means $\boldsymbol{\mu}_{1,2} = \mp\frac{\Delta}{2}\mathbf{v}$ along a single unit direction $\mathbf{v}$ and common isotropic within-class covariance $\sigma^2\mathbf{I}_d$, the true total covariance is the *rank-one* perturbation

$$\mathbf{\Sigma}_T(\sigma^2, \beta) = \sigma^2\mathbf{I}_d + \beta\,\mathbf{v}\mathbf{v}^\top, \qquad \beta := \frac{\Delta^2}{4}, \tag{16}$$

with eigenvalues $\sigma^2 + \beta$ (once) and σ^2 ($d-1$ times). Since Δ only contributes to $\mathrm{tr}(\mathbf{\Sigma}_T)$ along a single direction, its effect on equation 14 is $O(1)$, not $O(d)$ as the isotropic model implies – the two models disagree by a factor of d, which is exactly the quantity Weak-SIGReg is designed to scale gracefully with. This discrepancy is what motivates re-deriving the entanglement result from scratch under the correct model, rather than reusing Proposition 6's proof technique verbatim.

Exact closed form. Using $\mathrm{tr}(\mathbf{\Sigma}_T) = d\sigma^2 + \beta$ and $\mathrm{tr}(\mathbf{\Sigma}_T^2) = (\sigma^2+\beta)^2 + (d-1)\sigma^4$ inside equation 14 and simplifying gives, with $A = \frac{m(m+1)}{d^2}$, $B = \frac{m}{d^2}$, $D = \frac{2m}{d}$, the *exact* (not approximate) identity

$$L(\beta) = \Big(\frac{A}{d} + B\Big)(d\sigma^2 + \beta)^2 - D(d\sigma^2 + \beta) + m + A\,\frac{d-1}{d}\,\beta^2. \tag{17}$$

Because $\beta = \Delta^2/4$, L is an exact even quartic polynomial in Δ – not merely even to leading order, as in the Strong-SIGReg case.

Proposition 7 (Local NC1/NC2 entanglement, Weak-SIGReg)**.** *For fixed σ^2, $\Delta = 0$ is a strict local maximum of $L(\Delta)$ iff $\sigma^2 < \alpha^\star = \frac{d}{d+m+1}$, and a strict local minimum iff $\sigma^2 \geq \alpha^\star$. At $\sigma^2 = \alpha^\star$ the minimum is degenerate: the quadratic term in Δ vanishes and $L(\Delta) = L(0) + (A+B)\,\frac{\Delta^4}{16} > 0$ for $\Delta \neq 0$, so $\Delta = 0$ remains a strict (quartic) local minimizer.*

Proof. Differentiating equation 17 at $\beta = 0$:

$$\left.\frac{dL}{d\beta}\right|_{\beta=0} = 2d\sigma^2\Big(\frac{A}{d} + B\Big) - D = \frac{2m}{d}\Big[\sigma^2\,\frac{d+m+1}{d} - 1\Big], \tag{18}$$

which is zero iff $\sigma^2 = \alpha^\star$, negative for $\sigma^2 < \alpha^\star$, positive for $\sigma^2 > \alpha^\star$. Since $\beta = \Delta^2/4$ increases with $|\Delta|$, a negative derivative at $\beta = 0$ makes $\Delta = 0$ a strict local maximum of $L(\Delta)$, and a positive one makes it a strict local minimum. Since $\beta = \Delta^2/4$ increases with $|\Delta|$ and L is an exact upward parabola in β with curvature $2(A+B) > 0$, a negative (resp. positive) slope at $\beta = 0$ makes $\Delta = 0$ a strict local maximum (resp. minimum). At the degenerate point $\sigma^2 = \alpha^\star$ the slope vanishes but the quadratic term in β remains $(A+B)\beta^2 = (A+B)\Delta^4/16 > 0$, preserving strict minimality at order four. □

Proposition 8 (No entanglement at the global optimum)**.** *Over $\{\mathbf{\Sigma}_T(\sigma^2, \beta) : \sigma^2 \geq 0, \beta \geq 0\}$, the unique global minimizer of L is $\beta^\dagger = 0$ ($\Delta^\dagger = 0$), $\sigma^{2\dagger} = \alpha^\star$.*

Proof. Let $T_1 = \mathrm{tr}(\mathbf{\Sigma}_T) = d\sigma^2 + \beta$ and decompose $\mathrm{tr}(\mathbf{\Sigma}_T^2) = T_1^2/d + \mathrm{Gap}$, with $\mathrm{Gap} \geq 0$ by Cauchy–Schwarz (equality iff $\mathbf{\Sigma}_T$ is isotropic). Direct computation gives the exact identity $\mathrm{Gap}(\beta) = \beta^2(d-1)/d$, zero iff $\beta = 0$ (for $d > 1$). Substituting,

$$L = \Big(\frac{A}{d} + B\Big)T_1^2 - DT_1 + m + A \cdot \mathrm{Gap}. \tag{19}$$

The first three terms depend on (σ^2, β) only through T_1 and are minimized at $T_1 = T_1^\star := d\alpha^\star$ (vertex of this convex quadratic, cf. equation 15); the last term is minimized independently at $\beta = 0$. The point $\beta = 0, \sigma^2 = \alpha^\star$ satisfies $T_1 = d\alpha^\star = T_1^\star$ and $\mathrm{Gap} = 0$ simultaneously, so it minimizes both terms at once. Uniqueness: $\mathrm{Gap} = 0 \Rightarrow \beta = 0 \Rightarrow \sigma^2 = \alpha^\star$ from $T_1 = T_1^\star$. □

A.3 Deferred proofs: maximum entropy and uniformity on the sphere

Proof of Proposition 1. The first inequality is the maximum-entropy property of the Gaussian at fixed covariance. For the second, AM–GM on the eigenvalues gives $\det \boldsymbol{\Sigma} \leq (\mathrm{Tr}\boldsymbol{\Sigma}/d)^d$, with equality iff $\boldsymbol{\Sigma} \propto \mathbf{I}$; combining with $\frac{1}{2}\log(2\pi e)^d$ yields the bound, with equality throughout iff $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\mu}, \frac{\mathrm{Tr}\boldsymbol{\Sigma}}{d}\mathbf{I})$. □

Proposition 9 (Uniformity on the sphere)**.** *Among distributions on $\mathbb{S}^{d-1}$, the uniform distribution uniquely maximizes the differential entropy w.r.t. the surface measure. If $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$, then $\mathbf{Z}/\|\mathbf{Z}\|_2$ is uniform on $\mathbb{S}^{d-1}$ and $\|\mathbf{Z}\|_2/\sqrt{d} \to 1$ in probability as $d \to \infty$.*

Proof. Among densities p w.r.t. the (finite) surface measure μ, $-\int p \log p \, d\mu$ is uniquely maximized by $p \equiv 1/\mu(\mathbb{S}^{d-1})$ (Gibbs' inequality). Rotational invariance of $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ gives uniformity of $\mathbf{Z}/\|\mathbf{Z}\|_2$; by the law of large numbers $\|\mathbf{Z}\|_2^2/d \to 1$ almost surely, hence $\|\mathbf{Z}\|_2/\sqrt{d} \to 1$ in probability. □

A.4 Pinned Weak-SIGReg

Proof of Proposition 4. On the pinned set $\mathrm{Tr}\,\boldsymbol{\Sigma} = d_\ell$, the first term of Eq. equation 8 vanishes, and $\|\boldsymbol{\Sigma}\|_F^2 = \|\boldsymbol{\Sigma} - \mathbf{I}\|_F^2 + 2\,\mathrm{Tr}\,\boldsymbol{\Sigma} - d_\ell = \|\boldsymbol{\Sigma} - \mathbf{I}\|_F^2 + d_\ell$, which gives the stated identity, uniquely minimized at $\boldsymbol{\Sigma} = \mathbf{I}$. For the two-class model with $\sigma^2 = 1 - \Delta^2/(4d_\ell)$, $\boldsymbol{\Sigma}_T - \mathbf{I}$ has eigenvalue $\delta(1-1/d_\ell)$ on $\mathbf{v}$ and $-\delta/d_\ell$ on the $d_\ell - 1$ orthogonal directions, with $\delta = \Delta^2/4$, hence $\|\boldsymbol{\Sigma}_T - \mathbf{I}\|_F^2 = \delta^2(1-1/d_\ell)^2 + (d_\ell - 1)\delta^2/d_\ell^2 = \delta^2(d_\ell - 1)/d_\ell = (d_\ell - 1)\Delta^4/(16\,d_\ell)$, strictly increasing in $|\Delta|$. □

A.5 Same-class pairing probability and the collapse barrier

For a given sample and i.i.d. uniform class labels, each of the $B - 1$ other batch elements shares its class with probability $1/K$, independently, so the probability of having at least one same-class partner is

$$\Pr(\geq 1 \text{ same-class partner}) = 1 - (1 - 1/K)^{B-1}, \tag{20}$$

which evaluates to 0.92 for $K = 100$, $B = 256$ and to 0.40 for $K = 1000$, $B = 512$. Two regimes follow for the nearest-neighbor estimator equation 3. If the nearest neighbor is a same-class partner and the class has collapsed, $\varepsilon_1 \to 0$ and the maximization gradient of $\log \varepsilon_1$ has magnitude $1/\varepsilon_1$: a diverging barrier against further collapse. If instead the nearest neighbor belongs to another class, $\widehat{H}_{\mathrm{NN}}$ only measures the geometry of the class means and is insensitive to within-class collapse. The estimator is therefore informative exactly in proportion to equation 20, i.e. when B/K is moderate, and progressively blind as K/B grows.

A.6 Extended discussion

Local threshold vs. conditional optimum. It is worth distinguishing the local statement (Prop. 7: for σ^2 *held fixed* below $\alpha^\star$, moving Δ away from 0 decreases L) from a joint minimization over both variables (Prop. 8: jointly, the minimum sits at $\Delta = 0$). For a fixed $\sigma^2 < \alpha^\star$, the loss $L(\beta)$ is an exact upward parabola in β with unconstrained vertex

$$\beta^* = \frac{D - 2d\sigma^2(A/d + B)}{2(A + B)}, \tag{21}$$

which is strictly positive precisely when $\sigma^2 < \alpha^\star$ (same threshold as Prop. 7, since $\beta^* > 0 \iff dL/d\beta|_{\beta=0} < 0$). For $m = d - 1$ (our default sketch size, App. A.7), this evaluates to $\beta^* = \frac{d}{d+1}(1 - 2\sigma^2)$, i.e., for large d, the *partial-minimizer* $\Delta^2/4$ for a fixed σ^2 is approximately $2(\alpha^\star - \sigma^2)$ – about twice what the naive isotropic curve $\sigma^2 + \Delta^2/4 = \alpha^\star$ of a direct analogy with Strong-SIGReg would suggest. This partial-minimizer is *not*, however, where gradient descent on the *joint* objective converges: since $\beta = 0, \sigma^2 = \alpha^\star$ is also feasible and strictly better (Prop. 8), a training trajectory that is free to move σ^2 as well as Δ is not trapped by the local ridge at $\Delta = 0$ for $\sigma^2 < \alpha^\star$ – it can, and at the optimum does, resolve it by letting $\sigma^2 \to \alpha^\star$ rather than by growing Δ.

Relation to the Strong-SIGReg case. Proposition 6 states its threshold as $\sigma^2 \in (0, 1)$; an analogous joint-optimization argument is not immediate there because full-CF matching does not reduce to a two-statistic $(\operatorname{tr}\Sigma,\ \operatorname{tr}\Sigma^2)$ functional the way the Gaussian sketch does. This is precisely why the covariance-only relaxation of Akbar (2026), despite controlling fewer moments, admits an exact, closed-form characterization of *both* its local pathology and its true fixed point – an analysis that would not be available for the full-CF loss and that we believe is of independent interest beyond the present application.

Numerical sanity check and the value of $\alpha^\star$ in practice. For $d = 512$ and our default sketch size $m = d - 1 = 511$ (App. A.7, $C_{\text{batch}} = 512$ classes), $\alpha^\star = d/(d + m + 1) = 512/1024 = 0.5$. The pathology of Prop. 7 is therefore active over the whole range $\sigma^2 \in (0, 0.5)$ at this configuration: the protective regime $\alpha^\star \to 1$ only occurs for $m \ll d$, and should not be assumed in general. This is precisely why pinning the scale (Sec. 3) rather than relying on a large $\alpha^\star$ is necessary. For $\sigma^2 = 0.3 < \alpha^\star$, Prop. 7 predicts a repulsive ridge at $\Delta = 0$; for $\sigma^2 = 0.7 > \alpha^\star$, an attractive one.

A.7 Full training objective

We further discuss PIER introduced in equation 10, with its three terms respectively controlling *scale*, *shape*, and *isotropy*. These three terms play complementary roles:

- **Scale** (λ_{scale}). It pins each diagonal entry of $\widehat{\mathbf{\Sigma}}_T$ to 1, hence $\operatorname{Tr}\mathbf{\Sigma}_T \approx d_\ell$. This removes the free-trace channel behind Propositions 2 and 3, and the shrinkage bias $\alpha^\star$, so that we operate in the regime of Proposition 4.
- **Isotropy** (λ_{sigreg}). On the pinned set, $\mathcal{L}^{\text{glob}}_{\text{Weak}}$ is an isotropy penalty that penalizes concentrating variance in a few (class-separating) directions. It is cheap since it only involves an $m \times m$ sketch, and it does not suffer from the q-NN estimator's bias.
- **Shape** (λ_{entropy}). $\widehat{H}_{\text{NN}}$ (equation 3) acts on ℓ_2-normalized features, hence exerts no radial force: scale is left entirely to λ_{scale}, and the stop-gradient $\boldsymbol{\sigma}_\ell$ only preconditions the cosine geometry. It drives the angular distribution to uniformity: the angular marginal of $\mathcal{N}(\mathbf{0}, \mathbf{I})$ (Prop. 9) and adds, via the $1/\varepsilon_1$ gradient on same-class pairs, an anti-collapse barrier that second-order penalties lack.

Together, the terms drive p_ℓ towards an isotropic Gaussian with controlled scale, which is the maximum-entropy configuration of Proposition 1, one that keeps within-class variability away from zero.

B Implementation Details

B.1 Experimental Setup

Datasets. We train our models on ImageNet-100 (Tian et al., 2020) and ImageNet-1K (Russakovsky et al., 2015), which serve as the in-distribution (ID) datasets. ImageNet-1K contains 1,000 object categories and provides a diverse large-scale benchmark for image classification, while ImageNet-100 is a 100-class subset that enables experiments under a more limited training distribution.

To evaluate the transferability and OOD generalization of the learned representations, we consider eight downstream datasets spanning substantially different visual domains, image characteristics, and levels of semantic granularity: NINCO (Bitterwolf et al., 2023), ImageNet-R (Hendrycks et al., 2021), CIFAR-100 (Krizhevsky et al., 2010), Oxford 102 Flowers (Nilsback & Zisserman, 2008), CUB-200 (Wah et al., 2011), FGVC-Aircraft (Maji et al., 2013), Oxford-IIIT Pets (Parkhi et al., 2012), and STL-10 (Coates et al., 2011). Rather than evaluating only on datasets that closely resemble ImageNet, this collection includes natural images, fine-grained recognition tasks, low-resolution images, and domains with different object categories and visual statistics.

NINCO (No ImageNet Class Objects) (Bitterwolf et al., 2023) is specifically designed to provide a clean OOD benchmark for ImageNet. It contains 5,879 images from 64 OOD classes that were selected to have no categorical overlap with ImageNet-1K.

ImageNet-R (Hendrycks et al., 2021) contains renditions of ImageNet classes produced using artistic and other non-photographic styles, thereby introducing a substantial distribution shift in visual appearance while preserving semantic correspondence with the training categories.

CIFAR-100 (Krizhevsky et al., 2010) provides a substantially different image regime from ImageNet, consisting of 100 object classes represented by low-resolution 32×32 images. It consequently tests whether the learned representations remain useful when both image resolution and visual statistics differ substantially from those encountered during ImageNet training.

The remaining benchmarks emphasize fine-grained recognition. *Oxford 102 Flowers* (Nilsback & Zisserman, 2008) contains 102 flower categories and approximately 8,000 images, requiring discrimination between visually similar flower species. *CUB-200-2011* (Wah et al., 2011) contains 200 bird species and 11,788 images, providing a fine-grained recognition task with substantial intra-class variation. *FGVC-Aircraft* (Maji et al., 2013) contains 10,200 images spanning 102 aircraft variants and is organized hierarchically into variants, families, and manufacturers, making it another challenging fine-grained domain.

Oxford-IIIT Pets (Parkhi et al., 2012) contains images of 37 cat and dog breeds, with substantial variation in pose, scale, and illumination. Finally, *STL-10* (Coates et al., 2011) contains 10 object categories and 13,000 labeled images at 96×96 resolution, providing an additional low-resolution natural-image benchmark with a different category distribution from ImageNet.

Together, these datasets probe complementary forms of distribution shift, including changes in semantic domain, visual appearance, image resolution, and fine-grained category structure. We use them as downstream evaluation tasks to assess whether the representations learned from ImageNet retain their discriminative utility beyond the training distribution.

DNN Architectures. We train a set of backbones for ImageNet classification: ResNet18, ResNet34, ViT-Tiny, and ViT-Small, allowing direct comparison with entropy-regularization (Harun et al., 2025), SIGReg (Balestriero & LeCun, 2025), VISReg (Wu et al., 2026), VICReg (Bardes et al., 2021) and Weak-SIGReg (Akbar, 2026) baseline under identical architectural conditions. Unlike architectural approaches that mitigate collapse by modifying normalization layers (Pan & Cao, 2023; Harun et al., 2025), we use a regularization-based strategy that leaves the architecture unchanged. Our method is architecture-agnostic and can therefore be applied across different architectures and collapse regimes, including ViTs and CNNs with instance-based normalization.

Metrics. We report NC1 and NC2 as defined in Zhu et al. (2021) and Zhou et al. (2022), computed at the layer of interest. NC1 (feature collapse) is our primary metric, as it directly measures the quantity that our regularizer is designed to control. We report NC2 as a complementary metric to empirically assess the behavior predicted by Propositions 2 and 3, as well as the expected coupling between NC1 and NC2.

In addition, we report the in-distribution (ID) training accuracy to assess the impact of our regularizer on classification performance relative to standard cross-entropy (CE) training. Finally, we report the average OOD error rate (%) across eight different datasets. This metric evaluates the transferability of the representations learned on the ID dataset to diverse OOD distributions.

Training Details. We follow the training protocol of Harun et al. (2025): models are trained on ImageNet-100 and ImageNet1k for 100 epochs with the AdamW optimizer, a batch size of 512, a cosine learning-rate schedule with a 5-epoch linear warmup, weight decay of 0.05, label smoothing of 0.1, and 224×224 inputs with random resized crop and horizontal flip augmentation. Consistent with Harun et al. (2025), we use a learning rate of 0.01 for ResNets, 0.0016 for ViT-Tiny, and 0.0008 for ViT-Small. For the SGD-based ablation investigating stronger representation collapse, we use a learning rate of 0.1, momentum of 0.9, and weight decay of 5×10^{-4} for ResNet architectures.

Numerical Stability. When our method and an ablation yield numerically close results, we assess the stability of the observed differences across random seeds. Specifically, we repeat the pretraining procedure using 3 independent random seeds and evaluate each resulting model on all 8 OOD datasets using linear probing. We report the mean and standard deviation across the 3 runs to quantify the variability of the results.

B.2 Implementation Details.

Architecture Framework. Our architecture consists of a backbone network $f_\theta : \mathcal{X} \to \mathbb{R}^{d_\ell}$ that produces the pooled representation $\mathbf{z}_\ell$, with f_θ including the global average pooling layer. The regularizer $\mathcal{R}$ (Sec. 3) is applied directly to $\mathbf{z}_\ell$ and contributes only an auxiliary loss term. It does not modify the representation itself; consequently, the forward pass uses $\tilde{\mathbf{z}} * \ell \equiv \mathbf{z} * \ell$.

Regularization. During training, we apply the regularizer to encoder representations after global average pooling. The objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{CE} + \beta\mathcal{R},$$

where $\mathcal{R}$ is the chosen regularizer (Entropy, SIGReg, Weak SIGReg, or PIER). β balances classification accuracy and representation shaping; we fix $\beta = 0.05$. For the scale, entropy, and isotropic regularizers, we set the weight to 1 so they contribute equally. Although further tuning may improve results, we keep the setup simple and limit hyperparameters.

Feature Extraction for Linear Probing. We evaluate the learned representations through linear probing following the protocol of Harun et al. (2025); Papyan et al. (2020); Sarfi et al. (2023); Raghu et al. (2021). After pretraining, we freeze all network parameters and extract the next to last feature space z_ℓ. For each downstream dataset, we train a separate linear classifier on this frozen representation. For all architectures and downstream datasets, we use the same optimization protocol: learning rate 10^{-3}, weight decay 0, batch size 256, and 30 training epochs. We evaluate the resulting representations on eight OOD datasets.This protocol evaluates DNN transferability: the encoder features indicate representation quality and OOD robustness.generalization.

C Computational Cost

We express the computational cost of each regularization term in FLOPs, where B denotes the batch size, d the dimension of the embedding space, and m the number of sketch projections (for both Weak-SIGReg and our regularizer). A matrix multiplication $(n, k) \times (k, p)$ requires $2nkp$ FLOPs. We consider only the forward pass, as the backward-pass cost is proportional up to a constant factor (≈ 2).

The entropy term relies on computing pairwise similarities $zz^\top \in \mathbb{R}^{B\times B}$, which dominates its computational cost:

$$\text{FLOPs}_{\text{entropy}}(B, d) = 2B^2 d. \tag{22}$$

This cost is independent of m and grows linearly with the embedding dimension d.

Weak-SIGReg first projects the representations into an m-dimensional subspace using a sketching matrix $S \in \mathbb{R}^{m\times d}$. The projection $zS^\top$ requires $2Bdm$ FLOPs, after which the empirical covariance of the projected representations is computed at a cost of $2Bm^2$ FLOPs:

$$\text{FLOPs}_{\text{weak}}(B, d, m) = 2Bdm + 2Bm^2 = 2Bm(d + m). \tag{23}$$

Thus, the computational cost grows linearly with d and quadratically with m as m approaches d.

Our regularizer combines the entropy term with the Weak-SIGReg term. The additional centering and renormalization operations have complexit $O(Bd)$ and are therefore negligible compared with the two terms above:

$$\text{FLOPs}_{\text{PIER}}(B, d, m) = \text{FLOPs}_{\text{entropy}}(B, d) + \text{FLOPs}_{\text{weak}}(B, d, m) = 2B\left(Bd + dm + m^2\right). \tag{24}$$

The entropy term therefore introduces a fixed computational cost floor that is independent of m, while the marginal cost of the sketch remains approximately linear in m as long as $m \ll d$. As m approaches $d - 1$, however, the quadratic dependence on m becomes increasingly significant.

D ABLATION

D.1 FULL LOSS.

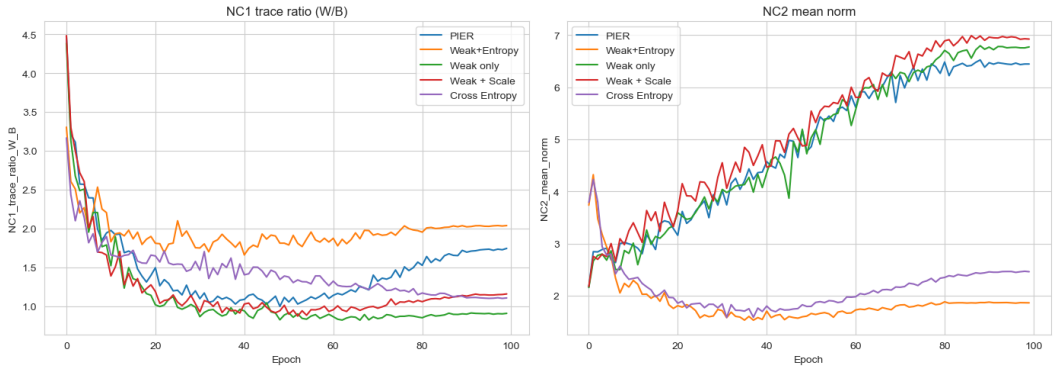

Figure 2: Evolution of representation collapse during ResNet18 training, with NC1 and NC2 reported throughout training.

Our objective combines three terms that respectively control the scale, isotropy, and shape of the representation (Eq. 10). We disentangle their contributions by ablating each component individually on ResNet-18 and ViT-Tiny. These two architectures allow us to assess both the necessity of scale control in architectures prone to collapse and its effectiveness in architectures that naturally maintain within-class variability.

Table 10 shows that WeakSIGReg alone does not prevent NC1 collapse, despite increasing NC2. The scale term addresses this issue by constraining the overall variance effectively preventing the loss to collapse bellow 1. As shown in Figure 2, NC1 remains pinned near 1 with scale. The entropy term is also important for ResNet-18 and performs better in combination with the other components than when used alone. Since the entropy objective only considers the nearest projected neighbors, it may push apart samples from different classes rather than specifically increasing separation among same-class samples. WeakSIGReg, however, naturally encourages different classes to occupy distinct projection directions. As training progresses and within-class separation increases, nearest neighbors are therefore increasingly likely to belong to the same class, making the entropy term progressively (epoch 60 in Figure 2) more effective.

For ViT-Tiny, which does not exhibit collapse, NC1 remains substantially higher throughout training, despite a gradual decrease. Consequently, explicit scale control is less critical in this setting. Nevertheless, constraining the scale during the early stages of training prevents the sharp reduction in within-class variance induced by cross-entropy, providing additional protection against collapse.

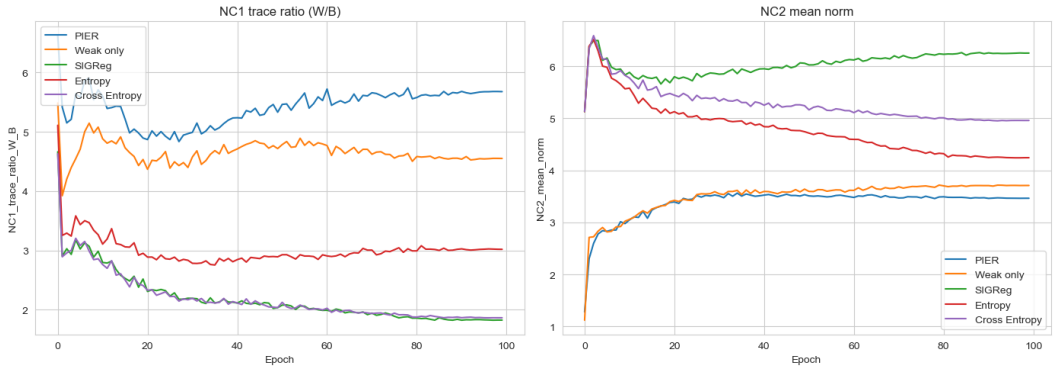

Figure 3: Evolution of representation collapse during ViT Tiny training, with NC1 and NC2 reported throughout training.

Since this constraint introduces negligible computational overhead and is particularly beneficial for architectures prone to collapse, we retain the scale term as a safeguard against the WeakSIGReg shortcut. As shown in Table 9, the scale and entropy terms complement each other, yielding better performance when combined than when entropy is applied to WeakSIGReg alone.

D.2 Effect of the Sketched Projection Size on ID and OOD Accuracy.

We evaluate the effect of the sketch projection size m and its relationship with $\alpha^\star$ described in Proposition 3 and further discussed in Appendix A.6. We conduct experiments on on ResNet-18 ($d = 512$) and ViT-Tiny ($d = 192$) while also reporting the corresponding number of FLOPs to quantify the computational cost associated with different projection sizes. We consider $m \in 128, 256, 352$ for ResNet-18, and $m \in 16, 64, 128$ for ViT-Tiny

As shown in Figure 4, for ResNet-18, our modification of WeakSGReg behaves as intended. Regardless of the projection size m, WeakSGReg consistently performs above the entropy baseline, suggesting that it does not fully address the first problem considered in our study. Our approach can combine WeakSGReg, pinned Scale, and entropy itself, while requiring a smaller projection dimension m. Consequently, for CNN architectures, our method can achieve a lower computational cost than WeakSGReg. Moreover, our method consistently performs below the entropy baseline in terms of the considered collapse-related metric, while achieving better transferability, suggesting improved OOD generalization compared with the current state of the art.

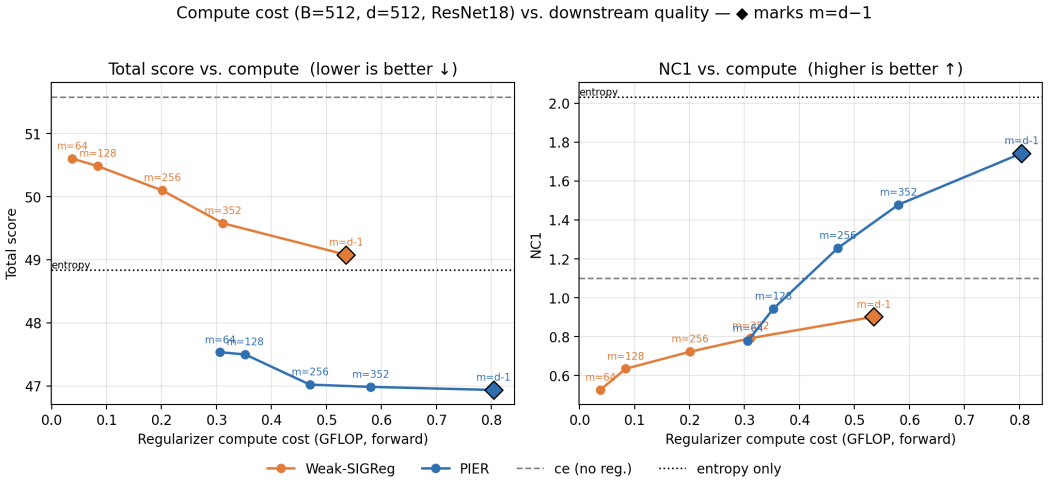

Figure 4: Score-to-computation and NC1-to-computation comparisons between Weak SIGReg and PIER using ResNet-18.

For ViT-Tiny, the embedding dimension d is relatively small, making the choice of m less critical. Overall, as Figure 5 shows, our method consistently achieves better anti-collapse behavior and, in this setting, also provides better transferability under matched m-to-m comparisons. In particular, using $m = d - 1$ yields the best accuracy with our method.

D.3 Effective Batch Size.

In the motivation Sec. 3, argue that the entropy estimator can be noisy for two reasons. First, its ability to detect within-class collapse depends on selecting a nearest neighbor belonging to the same class. This occurs with probability

$$1 - (1 - 1/K)^{B-1},$$

which decreases from 0.92 for $K = 100$, $B = 256$, to 0.40 for $K = 1000$, $B = 512$. Second, even when a same-class pair is selected, the resulting signal provides limited information about the remaining $d_\ell - 1$ directions of within-class variation.

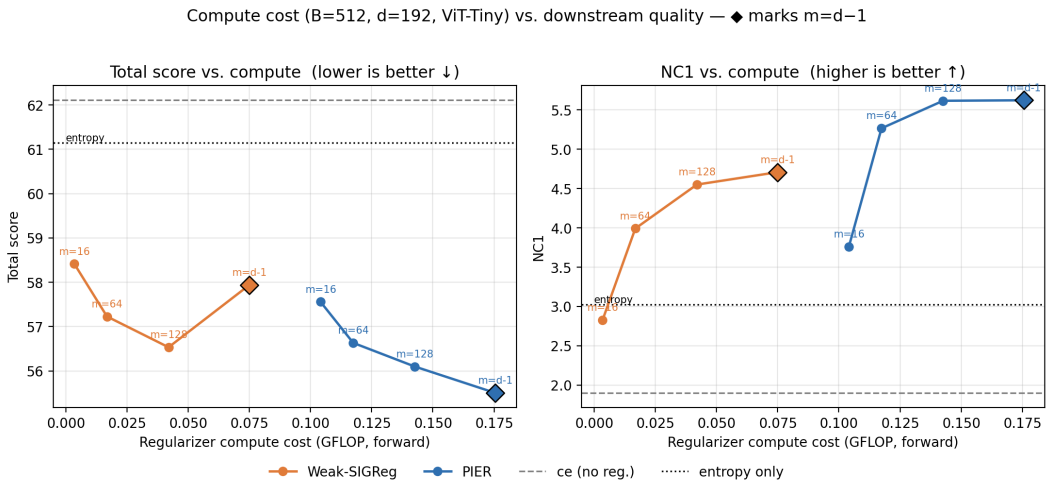

Figure 5: Score-to-computation and NC1-to-computation comparisons between Weak SIGReg and PIER using ViT-tiny.

The second aspect is evaluated throughout our experiments in the regime $K < B$, where we consistently observe that WeakSGReg and our proposed modification outperform the entropy baseline. However, when $K > B$, the behavior of entropy and WeakSGReg combined with our modification is less clear.

To investigate this regime, we evaluate ResNet18 and ViT-Tiny using batch sizes $B \in \{64, 128, 512\}$. The learning rate is scaled according to

$$\mathrm{LR}_{\mathrm{new}} = \mathrm{LR}_{\mathrm{ref}} \times \frac{B_{\mathrm{new}}}{512},$$

with $\mathrm{LR}_{\mathrm{ref}} = 0.01$ for ResNet18 and $\mathrm{LR}_{\mathrm{ref}} = 0.0016$ for ViT-Tiny.

For ResNet-18, we observe that the entropy baseline gradually improves as the batch size increases in the regime where the number of classes exceeds the batch size. In contrast, for $B = 64$ and $B = 128$, increasing the batch size does not lead to a substantial improvement.

For our method, substantial improvements are already observed at $B = 128$, where it becomes competitive with, and eventually outperforms, the entropy baseline. At $B = 512$, our method provides further improvements, achieving stronger performance than the compared SOTA approach.

Table 11: Error (%) as a function of batch size (*bs*) for different methods. Lower is better, mean over the 8 OOD datasets.

ResNet18

BS	Entropy	Weak d-1	PIER d-1
64	50.4425	50.5850	**49.2425**
128	49.8725	49.4913	**47.9625**
512	48.8375	49.0788	**46.9350**

ViT-Tiny

BS	Entropy	Weak	PIER
64	**62.9800**	63.1500	63.0063
128	62.9463	61.0050	**59.0288**
512	61.1440	57.9350	**55.5000**

D.4 More data augmentation.

The baseline uses limited data augmentation, namely RandomHorizontalFlip and RandomResizedCrop. For ViTs, however, stronger augmentation improves accuracy and becomes increasingly important at larger model scales, such as ViT-Small, to mitigate overfitting. We therefore investigate whether the improved NC metrics induced by stronger augmentation translate into better

OOD generalization when combined with our regularizer. Specifically, we use RandAugment with $num_{ops} = 2$ for both ViT-Tiny and ViT-Small.

The results in Tables 12 and 13 are consistent with our previous observations. Stronger augmentation further improves ViT-Tiny when combined with PIER, while ViT-Small reaches the performance ceiling achievable with WeakSigReg or PIER, with the latter providing the strongest regularization.

Table 12: ViT-Small on ImageNet-100 with RandAugment in addition to the default augmentation.

Model	$Acc_{\mathrm{ID}} \uparrow$	**Neural Collapse**		$\mathcal{E}_{\mathrm{GEN}}$ Avg. $\downarrow$
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
CE	69.12	1.39	7.13	54.37
Entropy	69.24	2.51	5.87	52.10
SIGReg	68.62	1.27	9.24	54.31
Weak-Sigreg	70.70	5.63	**4.44**	**46.70**
PIER	**71.22**	**6.90**	4.20	**46.48**

Table 13: ViT-Tiny on ImageNet-100 with RandAugment in addition to the default augmentation.

Model	$Acc_{\mathrm{ID}} \uparrow$	**Neural Collapse**		$\mathcal{E}_{\mathrm{GEN}}$ Avg. $\downarrow$
		$\mathcal{NC}_1$	$\mathcal{NC}_2$	
CE	69.56	1.63	5.37	59.51
Entropy	68.44	2.54	4.72	54.74
SIGReg	69.28	1.59	6.64	55.24
Weak-Sigreg	69.94	3.82	**4.04**	51.69
PIER	69.72	**4.53**	3.88	**46.01**

D.5 Increasing the Number of Classes.

Table 14 extends our analysis to $K = 1000$. At this scale, Entropy and SIGReg do not consistently improve over the CE baseline: they outperform it on only 2 of 5 backbones and perform up to 3.7% worse on the others. SIGReg remains particularly close to CE, with ID accuracy within 0.5% and NC1 within 3.7% across all five backbones, while NC2 typically increases by 9–10%. This behavior is consistent with the shortcut identified in Proposition 2: the regularizer largely follows the dynamics of CE without substantially changing the representation geometry, and the resulting increase in class separation does not consistently improve OOD generalization.

Weak-SIGReg and PIER. Weak-SIGReg and our proposed PIER are the only methods that consistently outperform CE across all five backbones. Both achieve the best $\mathcal{E}_{\mathrm{GEN}}$ in every setting, with a difference of only 0.01–0.15 points in most cases. The difference is larger under random augmentation, where our method achieves $\mathcal{E}_{\mathrm{GEN}} = 27.26$ compared to 27.90 for Weak-SIGReg.

Limits of anti-collapse regularization. The gains of our method over CE are smaller at $K = 1000$ (1.2–5.2%) than in the $K = 100$ setting of Sec. 4 (9–10.6%), and come with a small decrease in ID accuracy on ResNet-18 and ResNet-34 (-0.9 and -0.3 points, respectively). On ViT-Small, comparing models with and without random augmentation further illustrates the complementary role of the two approaches: augmentation alone reduces $\mathcal{E}_{\mathrm{GEN}}$ by 11.9%, compared with 1.2% from anti-collapse regularization alone. When combined, our method still provides an additional 5.2% improvement over augmentation. Thus, at ImageNet-1K scale, anti-collapse regularization remains effective, but its gains are more limited when used alone.

E More Results.

In this section, we provide the complete results for our ImageNet-100 experiments (Table 15), extending the comparisons with VISReg and VICReg alongside the regularizers evaluated in the main paper. We also report the full results of our ablations on projector size (Tables 16 and 17) and label smoothing (Table 18).

Table 14: Summary of the ImageNet-1K results. The table reports ID accuracy, the first two neural-collapse metrics, and the average OOD generalization error as mean ± standard deviation. The best OOD generalization error for each backbone is shown in bold.

Backbone	**Model**	$Acc_{\mathrm{ID}} \uparrow$	$\mathcal{NC}_1 \uparrow$	$\mathcal{NC}_2 \uparrow$	$\mathcal{E}_{\mathrm{GEN}}$ Avg. $\downarrow$
ResNet18	CE	70.22	1.93	5.78	28.71 ± 0.16
	Entropy	70.04	2.09	5.53	28.93 ± 0.06
	SIGReg	70.22	1.86	6.30	28.89 ± 0.07
	Weak	69.80	2.39	6.43	$\mathbf{27.78 \pm 0.16}$
	PIER	69.59	2.57	6.45	27.89 ± 0.07
ResNet34	CE	72.57	1.71	5.69	29.29 ± 0.16
	Entropy	72.43	1.85	5.50	28.98 ± 0.09
	SIGReg	72.27	1.66	6.28	29.00 ± 0.09
	Weak	72.48	2.00	6.91	28.45 ± 0.21
	PIER	72.33	2.18	6.88	$\mathbf{28.34 \pm 0.16}$
ViT-Tiny	CE	67.92	1.63	10.16	32.38 ± 0.12
	Entropy	68.44	1.87	9.88	33.36 ± 0.27
	SIGReg	67.95	1.69	9.80	33.57 ± 0.10
	Weak	68.84	1.72	5.68	31.05 ± 0.11
	PIER	**69.99**	1.79	5.67	$\mathbf{31.04 \pm 0.11}$
ViT-Small	CE	66.31	2.66	8.26	32.61 ± 0.03
	Entropy	66.43	3.27	7.88	33.56 ± 0.10
	SIGReg	66.18	2.65	9.09	33.71 ± 0.14
	Weak	66.19	2.83	6.01	32.34 ± 0.13
	PIER	66.41	3.19	5.84	$\mathbf{32.21 \pm 0.21}$
ViT-Small with random data augmentation	CE	71.28	2.37	8.83	28.74 ± 0.16
	Entropy	71.22	2.89	8.49	28.10 ± 0.04
	SIGReg	70.91	2.39	9.66	28.72 ± 0.07
	Weak	71.92	2.52	6.35	27.90 ± 0.18
	PIER	**72.26**	2.75	6.22	$\mathbf{27.26 \pm 0.10}$

Table 15: ImageNet-100 results across the eight OOD datasets. The table reports ID accuracy, the first two neural-collapse metrics, individual OOD errors, and their average. Lower values are better for the OOD metrics.

Backbone	**Model**	$Acc_{\mathrm{ID}}\uparrow$	$\mathcal{NC}_1\uparrow$	$\mathcal{NC}_2\uparrow$	imnet_r200	cifar	flower	ninco	cub	aircrafts	pet	stl	$\mathcal{E}_{\mathrm{GEN}}$ Avg. $\downarrow$
ResNet18	CE	83.12	1.10	2.47	77.78	49.61	43.92	32.14	78.22	83.05	29.98	17.89	51.57
	Entropy	83.10	2.03	1.86	74.77	46.53	39.02	29.08	73.39	83.05	28.35	16.51	48.84
	SigReg	83.76	1.03	2.68	77.30	49.15	42.35	31.46	77.06	82.81	32.32	17.49	51.24
	VisReg	83.00	0.77	1.53	77.98	51.15	45.10	34.44	78.55	88.96	32.41	18.59	53.40
	VICReg	83.08	1.04	2.82	76.13	49.09	45.88	30.10	76.58	82.99	30.72	17.31	51.10
	Weak	83.08	0.90	**6.76**	75.88	46.98	40.98	28.83	74.47	79.87	28.92	16.70	49.08
	PIER	82.86	1.74	6.45	74.60	45.79	36.57	26.70	70.21	78.40	27.50	15.71	**46.94**
ResNet34	CE	84.36	0.90	2.28	77.68	48.90	50.59	34.61	81.34	89.23	35.13	18.52	54.50
	Entropy	84.26	1.89	1.79	75.25	46.96	43.73	30.44	76.68	86.38	31.34	16.83	50.95
	SigReg	84.52	0.86	2.51	77.72	49.81	46.08	34.61	82.17	88.87	34.81	17.79	53.98
	VisReg	84.24	0.55	1.74	79.72	53.92	55.59	37.50	83.91	92.26	37.75	20.79	57.68
	VICReg	84.30	0.84	2.73	77.47	49.74	48.33	33.67	81.93	88.48	33.88	17.54	53.88
	Weak	84.44	0.68	**7.71**	76.17	48.33	46.76	30.10	80.50	86.47	33.93	16.87	52.39
	PIER	84.36	1.52	7.00	74.80	46.07	36.37	26.19	74.08	83.32	29.44	14.91	**48.15**
ViT-Tiny	CE	62.86	1.90	5.01	86.62	56.46	59.51	43.11	79.24	86.98	55.46	29.46	62.11
	Entropy	61.46	3.02	4.29	85.95	56.34	55.88	40.05	78.08	86.50	55.60	30.75	61.14
	SigReg	62.46	1.82	6.28	86.42	56.22	58.24	41.75	80.27	86.50	55.90	30.55	61.98
	VisReg	61.18	1.98	4.09	87.40	58.63	63.33	42.35	80.58	88.15	55.85	30.31	63.33
	VICReg	62.68	3.16	4.27	85.35	53.18	50.69	36.56	75.85	85.42	53.34	27.72	58.51
	Weak	62.76	4.71	**3.67**	85.72	53.62	47.65	36.73	74.09	85.42	52.06	28.19	57.94
	PIER	**63.74**	**5.62**	3.48	83.78	52.99	42.45	34.95	70.76	83.17	49.74	26.17	**55.50**
ViT-Small	CE	62.46	1.59	6.60	85.22	53.40	56.86	41.67	79.44	86.59	54.48	29.89	60.94
	Entropy	62.60	3.06	5.33	83.53	51.11	45.59	36.65	73.18	80.02	50.23	26.91	55.90
	SigReg	62.02	1.50	8.91	84.65	52.67	56.08	40.48	79.05	85.24	54.18	28.97	60.17
	VisReg	60.56	2.61	4.57	83.57	50.72	49.41	36.56	75.23	84.82	51.65	27.31	57.41
	VICReg	63.98	1.83	4.17	85.58	56.37	59.12	44.56	80.84	87.43	57.10	30.95	62.74
	Weak	63.98	7.73	**4.07**	81.03	48.84	37.35	32.91	67.92	78.79	46.12	24.19	**52.14**
	PIER	64.20	**9.32**	3.81	82.28	48.65	37.75	31.63	67.28	78.40	47.53	25.49	**52.38**

Table 16: Ablation study of the projection dimension for ResNet18 across the eight OOD datasets. The table reports the first two neural-collapse metrics, individual OOD errors, and their average. Lower values are better.

Model	m	$\mathcal{NC}_1\uparrow$	$\mathcal{NC}_2\uparrow$	imnet_r200	cifar	flower	ninco	cub	aircrafts	pet	stl	$\mathcal{E}_{\mathrm{GEN}}$ Avg. $\downarrow$
CE	–	1.10	2.47	77.78	49.61	43.92	32.14	78.22	83.05	29.98	17.89	51.57
Entropy	–	2.03	1.86	74.77	46.53	39.02	29.08	73.39	83.05	28.35	16.51	48.84
Weak	64	0.52	9.64	75.97	47.34	44.41	29.51	76.56	82.87	31.02	17.14	50.60
Weak	128	0.63	9.88	75.62	48.95	42.75	31.63	75.94	81.79	30.12	17.06	50.48
Weak	256	0.72	8.51	76.40	48.26	42.84	30.61	74.20	82.12	29.60	16.75	50.10
Weak	352	0.79	7.84	76.38	47.90	40.78	29.17	74.47	80.38	30.96	16.58	49.58
PIER	64	0.78	9.93	74.37	45.82	37.65	27.21	71.26	79.99	28.81	15.16	47.53
PIER	128	0.94	10.05	73.07	47.04	36.76	26.28	71.82	80.14	28.81	16.06	47.50
PIER	256	1.25	2.22	73.73	46.30	34.12	27.72	70.85	79.39	28.48	15.57	47.02
PIER	**352**	1.48	7.47	74.50	45.67	34.51	27.04	72.30	78.58	28.05	15.22	**46.98**

Table 17: Ablation study of the projection dimension for ViT-Tiny across the eight OOD datasets. The table reports the first two NC metrics, individual OOD errors, and their average. Lower values are better.

Model	m	$\mathcal{NC}_1 \uparrow$	$\mathcal{NC}_2 \uparrow$	imnet_r200	cifar	flower	ninco	cub	aircrafts	pet	stl	$\mathcal{E}_{\text{GEN}}$ Avg. $\downarrow$
CE	–	1.90	5.01	86.62	56.46	59.51	43.11	79.24	86.98	55.46	29.46	62.11
Entropy	–	3.02	4.29	85.95	56.34	55.88	40.05	78.08	86.50	55.60	30.75	61.14
Weak	16	2.83	6.38	85.45	54.58	48.53	37.84	76.35	84.43	52.33	27.80	58.41
Weak	64	3.99	4.95	84.30	53.95	48.33	35.54	73.52	83.89	50.56	27.72	57.23
Weak	128	4.55	4.10	84.85	52.94	46.67	33.50	72.21	84.07	51.27	26.75	56.53
PIER	16	3.76	5.81	84.53	53.63	47.94	37.41	73.51	83.95	52.03	27.46	57.56
PIER	64	5.27	4.50	84.23	53.68	45.20	35.03	72.75	83.56	51.19	27.41	56.63
PIER	**128**	**5.62**	3.86	84.47	53.70	44.51	34.18	71.82	82.33	50.67	27.10	**56.10**

Table 18: Results with label smoothing $= 0.0$ across the eight OOD datasets. The table reports ID accuracy, the first two NC metrics, individual OOD errors, and their average. Lower values are better for the OOD metrics.

Backbone	**Model**	$Acc_{\text{ID}} \uparrow$	$\mathcal{NC}_1 \uparrow$	$\mathcal{NC}_2 \uparrow$	imnet_r200	cifar	flower	ninco	cub	aircrafts	pet	stl	$\mathcal{E}_{\text{GEN}}$ Avg. $\downarrow$
ResNet18	Entropy	81.82	1.91	4.40	73.18	44.86	32.84	26.36	67.22	76.45	27.01	15.20	45.39
	Weak	81.82	2.21	5.48	71.32	45.31	32.84	24.66	66.64	75.76	25.59	14.89	44.63
	PIER	**82.60**	2.91	5.44	71.62	42.69	29.80	25.17	65.69	75.04	25.10	14.75	**43.73**
ResNet34	Entropy	82.64	1.87	4.17	72.28	44.75	35.29	26.87	67.64	78.25	27.83	15.21	46.02
	Weak	**83.22**	2.03	5.76	70.77	43.87	31.96	25.51	66.43	76.45	26.14	14.44	44.45
	PIER	82.66	2.69	5.59	71.05	43.57	29.51	22.70	67.29	76.03	25.51	13.96	**43.70**
ViT-Tiny	Entropy	61.14	4.42	7.28	84.48	52.28	46.27	34.35	73.42	83.14	51.95	23.35	56.16
	Weak	61.62	5.88	3.54	83.82	52.42	45.98	34.18	70.23	82.60	51.13	26.79	55.89
	PIER	**62.76**	**6.06**	3.58	82.93	51.11	44.41	33.42	70.62	82.18	51.10	26.45	**55.28**