Class-Conditional Activation Regularization (CCAR): Intrinsic Robustness as an Emergent Geometric Property

Akash Samanta [*1] **Manish Pratap Singh** [*2] **Debasis Chaudhuri** [*1]

Abstract

Standard supervised learning optimizes for predictive accuracy but remains agnostic to the internal geometry of learned features, often yielding representations that are entangled and brittle. We propose Class-Conditional Activation Regularization (CCAR) to explicitly engineer the feature space, imposing a block-diagonal structure via a soft inductive bias. By shaping the latent representation to confine class energy to orthogonal subspaces, we create an intrinsic geometric scaffold that naturally filters noise and single-step adversarial perturbations. We provide theoretical analysis linking this structural constraint to the maximization of the Fisher Discriminant Ratio, establishing a formal connection between geometric disentanglement and algorithmic stability. Empirically, this approach demonstrates that robustness is an emergent property of a well-engineered feature space, significantly outperforming baselines on label noise and input corruption benchmarks.

1. Introduction

The remarkable success of deep neural networks stems from their ability to discover hierarchical representations that disentangle complex data manifolds (Bengio et al., 2013). Yet, standard supervised training focuses predominantly on optimizing prediction error. While effective at minimizing empirical risk, this approach remains largely agnostic to internal feature structure. Consequently, networks often learn degenerate geometries sufficient for clean data but brittle under stress, collapsing when exposed to label noise (Zhang et al., 2017), adversarial perturbations, or distribution shifts (Geirhos et al., 2020).

[*]Equal contribution [1]Techno India University, Salt Lake, Kolkata, India [2]DRDO Young Scientist Laboratory - CT, Chennai, India. Correspondence to: Akash Samanta <akash-samanta.web@gmail.com>.

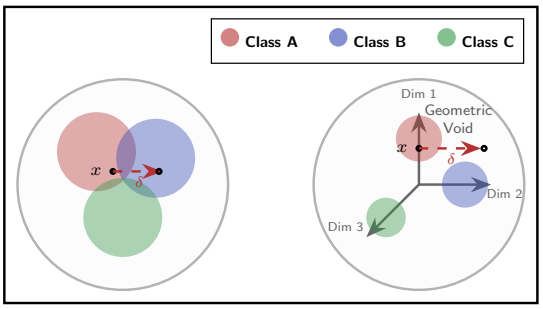
Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

This disconnect raises a fundamental question: can we explicitly shape the internal geometry of deep representations to be intrinsically robust, without relying on external augmentations or complex training pipelines?

Current robustness methods typically intervene at the periphery via data augmentations (Cubuk et al., 2019), adversarial training (Madry et al., 2018), or specialized architectural modules. While effective, they often treat the symptoms of fragile geometry rather than addressing it as a critical structural vulnerability. Because geometric entanglement leaves decision boundaries dangerously close to the data manifold, we posit that explicitly enforcing orthogonal geometry acts as a causal, mechanistic defense. This demonstrates that robustness is an emergent property of well-structured feature spaces (Ilyas et al., 2019). If latent representations are organized into distinct, stable subspaces aligned with semantic identity, the network naturally resists perturbations without needing to encounter every possible noise pattern during training.

In this work, we propose Class-Conditional Activation Regularization (CCAR), a versatile, readily integrable geometric regularizer for supervised representation learning designed to bridge the gap between predictive optimization and structural robustness. We explicitly differentiate this approach from highly engineered noisy-label pipelines (e.g., DivideMix (Li et al., 2020)) or dedicated adversarial training frameworks that require multi-stage or computationally expensive interventions. Unlike these bespoke systems, or standard regularizers that merely constrain model complexity, CCAR acts as a passive, soft inductive bias directly on the intermediate activations during standard training. By designating soft, contiguous subspaces for each class within the high-dimensional feature manifold, we encourage the network to concentrate feature energy into class-specific directions. Crucially, while this enforces class-specific partitions, it does not enforce rigid sparsity; neurons remain polysemantic within their assigned subspaces, and the aggregate geometry becomes significantly more coherent.

Our approach draws inspiration from the principle that semantic separation should exist not just at the decision boundary, but throughout the latent representation. We demonstrate that a simple geometric constraint, implemented as



(a) Standard Supervision (b) CCAR (Ours)

Figure 1. Geometric intuition of emergent robustness. (Left) Standard Supervision: Training typically induces dense, entangled class manifolds with shared boundaries. A small perturbation (δ) easily pushes a sample x (Class A) into a competing manifold (Class B). (Right) CCAR (Ours): By enforcing a block-diagonal structure, classes are confined to orthogonal subspaces. This creates a large geometric void (inactive feature space) between manifolds. The same perturbation δ now projects into this void rather than a competing class, preserving the correct prediction.

a smooth quadratic penalty accompanying the primary supervision, is sufficient to induce this structure. This complementary regularization forces the network to learn features that are not only discriminative but also compact and centered, reducing the manifold instability where adversarial examples and noise typically hide.

We summarize our primary contributions as follows:

- **General-Purpose Geometric Regularizer:** We introduce CCAR, a lightweight, plug-in framework for supervised representation learning. By directly shaping intermediate feature geometry, it bypasses costly data-cleaning or adversarial generation pipelines, significantly improving cluster separation and reducing feature overlap.
- **Robustness via Structure:** We demonstrate that shaping geometry yields emergent robustness. CCAR outperforms standard baselines on label noise, input corruptions, and single-step adversarial attacks (FGSM), proving robustness arises from organized internal representations rather than mere boundary optimization.
- **Theoretical & Empirical Analysis:** We theoretically characterize CCAR as an energy minimization problem under class constraints, and empirically show that geometric improvements (e.g., maximizing Fisher Discriminant Ratio) correlate with downstream stability.

2. Background and Related Work

Robust representation learning seeks an encoder $f_\theta : \mathcal{X} \to \mathbb{R}^D$ whose feature geometry remains stable under perturbations to the input $P(X)$ or label $P(Y)$ distributions. Standard supervised learning optimizes predictive accuracy but ignores the geometry of the latent space $\mathbb{R}^D$, yielding separable yet geometrically fragile representations. We review the geometric properties of standard losses, invariance limitations of contrastive objectives, and existing robustness mechanisms.

2.1. Feature Geometry in Supervised Learning

In the standard supervised classification setting, a network learns a mapping $f_\theta(x)$ and a linear classifier $W = [w_1, \dots, w_C]^\top$ to minimize the Cross-Entropy (CE) loss (Shannon, 1948):

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(w_{y_i}^\top f_\theta(x_i))}{\sum_{j=1}^{C} \exp(w_j^\top f_\theta(x_i))} \tag{1}$$

Analytically, minimizing $\mathcal{L}_{\text{CE}}$ aligns $f_\theta(x_i)$ with target w_{y_i} while pushing inner products with $w_{j \neq y_i}$ toward a symmetric negative equilibrium (Simplex ETF). However, this optimization is under-determined regarding feature magnitude and energy distribution. The encoder freely utilizes arbitrary subspaces or highly correlated neurons, inducing feature collapse—compressing intra-class variability only along training-specific directions—which leaves representations vulnerable to out-of-distribution noise or adversarial shifts (Papyan et al., 2020).

Auxiliary objectives like Center Loss (Wen et al., 2016) and ArcFace (Deng et al., 2019) regularize this geometry by enforcing relative metric constraints (e.g., minimizing centroid distances or enforcing angular margins). While effective, they constrain relative point distances rather than absolute feature activation patterns.

2.2. Contrastive Learning and Geometric Invariance

Contrastive Learning (CL) creates structure by aligning positive pairs (x, x^+) and repelling negatives x^-, typically optimized via objectives like the InfoNCE loss. However, a critical limitation of CL for structured feature learning is its invariance to orthogonal transformations. Because these objectives depend solely on pairwise inner products or Euclidean distances, the optimal latent geometry is defined only up to an arbitrary rotation.

Proposition 2.1 (Rotation Symmetry)**.** *For any loss $\mathcal{L}$ depending only on pairwise inner products $\langle f(x), f(x') \rangle$, if f^* is a minimizer, then Rf^* is also a minimizer for any orthogonal matrix $R \in O(D)$.*

This symmetry implies that CL cannot enforce axis-aligned disentanglement without additional constraints. While Non-negative Contrastive Learning (NCL) (Wang et al., 2024) breaks this symmetry via non-negativity to induce

sparse, axis-aligned features, it operates strictly in the self-supervised regime. Furthermore, even Supervised Contrastive Learning (SupCon) (Khosla et al., 2020) inherits this rotational invariance despite utilizing label information. Consequently, although CL improves global separability, it cannot guarantee that specific classes occupy stable, distinct coordinate subspaces.

2.3. Robustness via Inductive Biases

Approaches to enforce robustness typically intervene either at the data level or the optimization level. Data-level methods, such as Adversarial Training, solve a min-max problem against a perturbation set $\mathcal{S}$:

$$\min_{\theta,W} \mathbb{E}\left[\max_{\delta\in\mathcal{S}} \mathcal{L}(W f_\theta(x+\delta), y)\right] \tag{2}$$

This explicitly constructs the decision boundary around worst-case noise but is computationally intensive (Madry et al., 2018). Optimization-level methods, such as DivideMix (Li et al., 2020) or robust loss functions (Ghosh et al., 2017), attempt to filter or down-weight noisy samples.

A third paradigm, geometric regularization, posits structural disentanglement as a mechanistic defense against feature fragility. While methods like Maximal Coding Rate Reduction (MCR^2) and Orthogonal Projection Loss encourage orthogonal class subspaces, they require computationally heavy covariance matrix tracking or eigendecompositions. In contrast, CCAR is a seamlessly integrable, general-purpose enhancement that uses an axis-aligned, activation-space soft L_2 penalty. By enforcing strict partitions directly on individual activations without requiring class centroids or complex matrix operations, CCAR explicitly penalizes feature leakage into forbidden regions to guarantee robustness, distinguishing it from general covariance-based methods (Lezama et al., 2018).

3. Class-Conditional Geometric Regularization

Standard supervised learning paradigm optimizes an encoder $f_\theta : \mathcal{X} \to \mathbb{R}^D$ and a linear classifier W to minimize the cross-entropy loss. While this objective successfully maximizes the separation between class centroids, it remains agnostic to the internal geometry of the feature space $\mathbb{R}^D$. Consequently, the representations often exhibit high variance in directions parallel to the decision boundary (orthogonal to its normal); these act as nuisance directions that contribute nothing to classification but serve as avenues for adversarial perturbations (Geirhos et al., 2020; Ilyas et al., 2019). This unchecked entanglement constitutes a critical structural vulnerability, leaving decision boundaries in close proximity to the training manifold.

Algorithm 1 CCAR-L2 Supervised Training

Require: Batch $\mathcal{B} = \{(x_i, y_i)\}_{i=1}^{|\mathcal{B}|}$,
Encoder f_θ, Classifier W, Strength λ
Forward Pass:
$H \leftarrow f_\theta(X)$
$Z \leftarrow WH$
$\mathcal{L}_{\text{CE}} \leftarrow \text{CrossEntropy}(Z, Y)$
Geometric Regularization:
Initialize $\mathcal{L}_{\text{reg}} \leftarrow 0$
$c_{\text{norm}} \leftarrow |\mathcal{O}_y| = K(C-1)$ *// Constant for a given configuration*
for $i = 1$ **to** $|\mathcal{B}|$ **do**
  $m_i \leftarrow \text{Mask}(y_i)$
  $E_i \leftarrow \|m_i \odot H_i\|_2^2$
  $\mathcal{L}_{\text{reg}} \leftarrow \mathcal{L}_{\text{reg}} + E_i$
end for
$\mathcal{L}_{\text{reg}} \leftarrow \frac{1}{c_{\text{norm}}}\mathcal{L}_{\text{reg}}$
Backward Pass:
$\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{CE}} + \frac{\lambda}{|\mathcal{B}|}\mathcal{L}_{\text{reg}}$
Update θ, W via SGD using $\nabla\mathcal{L}_{\text{total}}$

To remedy this, we introduce Class-Conditional Activation Regularization (CCAR), a geometric framework that explicitly imposes a block-diagonal covariance structure on latent representations.

3.1. Subspace Partitioning

We define robust latent geometry by confining class information to orthogonal subspaces, transforming entangled manifolds into structures pinned to designated axes (Figure 1). Specifically, we partition D into C disjoint sets $\mathcal{P} = \{\mathcal{I}_1, ..., \mathcal{I}_C\}$ where $\mathcal{I}_c = \{(c-1)K+1, \ldots, cK\}$. To balance capacity, we set $|\mathcal{I}_c| = K = \lfloor D/C \rfloor$ and suppress any remainder dimensions $D \pmod C$. While this contiguous layout is a practical choice, random-fixed assignments yield equivalent robustness (Appendix C.3), proving that orthogonality rather than specific coordinate ordering drives these benefits.

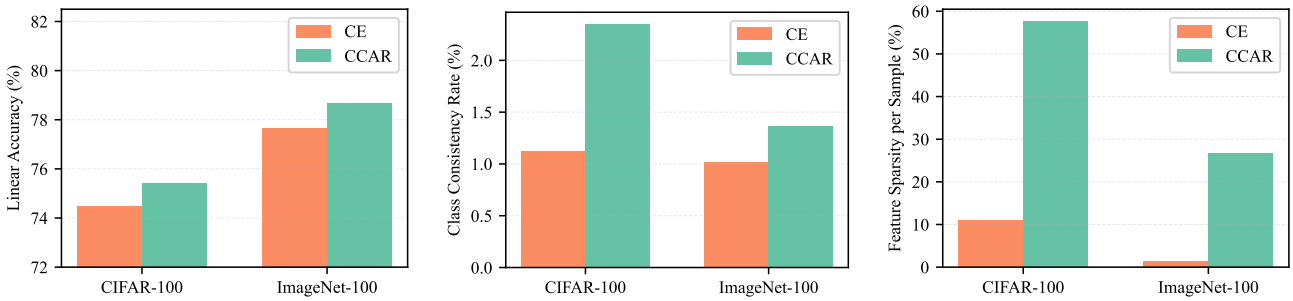
For an input x associated with label y, the ideal feature representation $h = f_\theta(x)$ should reside strictly within the subspace spanned by the basis vectors in $\mathcal{I}_y$. We define the *forbidden region* $\mathcal{O}_y$ as the orthogonal complement of the active set:

$$\mathcal{O}_y = \{1, \ldots, D\} \setminus \mathcal{I}_y \tag{3}$$

To operationalize this constraint, we define a class-conditional projection operator $M(y) \in \{0,1\}^D$ that isolates the forbidden dimensions:

$$(M(y))_j = \mathbb{I}(j \in \mathcal{O}_y) \tag{4}$$

Remark on Polysemanticity. It is important to distinguish



(a) Linear Accuracy (b) Consistency Rate (c) Feature Sparsity

Figure 2. Comparison between Cross-Entropy (CE) and Class-Conditional Activation Regularization (CCAR): (a) linear probing accuracy; (b) class consistency rate (absolute %), measuring the proportion of top-10 maximally activating samples belonging to a neuron's assigned class; (c) feature sparsity, the average proportion of zero elements ($|x| < 1\text{e-}5$) in each test sample's features. CCAR significantly improves consistency and sparsity without sacrificing accuracy.

between *subspace constraints* and *neuron monosemanticity*. CCAR allocates a dedicated subspace $\mathcal{I}_y$ to each class, but it imposes no sparsity constraints *within* that subspace. The network retains the freedom to learn distributed, polysemantic representations, where single neurons encode multiple intra-class attributes (Olah et al., 2020; Elhage et al., 2022), within the active block. This structure preserves the coding efficiency of high-dimensional spaces while eliminating cross-class interference.

3.2. The CCAR-L2 Objective

We formulate the regularization as an energy minimization problem on the forbidden projection. Let $h = f_\theta(x)$ be the feature representation. The geometric loss is defined as the mean squared energy of the residual energy into forbidden dimensions:

$$\mathcal{L}_{\text{CCAR}}(x, y) = \frac{1}{|\mathcal{O}_y|}\|M(y) \odot h\|_2^2 = \frac{1}{|\mathcal{O}_y|}\sum_{j \in \mathcal{O}_y} h_j^2 \quad (5)$$

We adopt the squared L_2 norm over sparsity-inducing alternatives (e.g., L_1) to ensure optimization stability. The L_2 penalty provides a Lipschitz continuous gradient field, offering a restorative force scaling linearly with deviation. This prevents the oscillatory convergence near the manifold origin characteristic of non-smooth penalties (Bertsekas, 1999), facilitating precise subspace convergence. Appendix C provides a comprehensive theoretical and empirical comparison justifying L_2 over alternative norms.

3.3. Optimization Algorithm

The final training objective is a composite loss that balances discriminative performance with geometric compression. We combine the standard Cross-Entropy loss ($\mathcal{L}_{\text{CE}}$) with the CCAR regularizer:

$$\min_{\theta, W} \mathcal{J} = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathcal{L}_{\text{CE}}(W f_\theta(x), y) + \lambda \mathcal{L}_{\text{CCAR}}(f_\theta(x), y)\right] \quad (6)$$

where λ is a scalar hyperparameter controlling the regularization strength. To ensure the practical feasibility of this objective, we characterize the optimization landscape under the L_2 penalty. Unlike non-smooth regularizers, the quadratic nature of our constraint ensures stable convergence behavior.

Computational Complexity. The overhead introduced by CCAR is negligible. The mask generation and element-wise multiplication are $O(D)$ operations, which can be fully parallelized on GPU hardware. Unlike contrastive methods that require $O(N^2)$ pairwise computations or costly negative mining (Chen et al., 2020), CCAR operates independently on each sample, scaling linearly $O(N)$ with batch size.

4. Theoretical Analysis

In this section, we analyze the geometric properties induced by the CCAR-L2 objective. We demonstrate that minimizing the forbidden energy is statistically equivalent to maximizing the Fisher Discriminant Ratio in the latent space. Furthermore, we demonstrate a geometric stability bound, proving that the block-diagonal structure creates a certified margin against additive perturbations.

4.1. Covariance Compression and Fisher Ratio

We first characterize the effect of CCAR on the statistical structure of the learned features. Let $\Sigma_c = \mathbb{E}[(h - \mu_c)(h - \mu_c)^\top | y = c]$ denote the class-conditional covariance matrix for class c. In unregularized networks, Σ_c is typically full-rank, indicating variance in all directions.

Theorem 4.1 (Trace Minimization)**.** *Assuming class means are aligned with their active subspaces ($P_{\mathcal{O}_c}\mu_c = 0$, where*

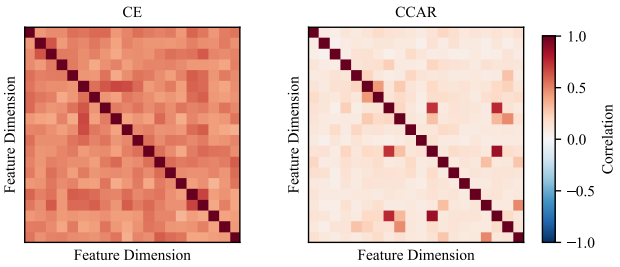


Figure 3. Dimensional correlation matrix of 20 random features on ImageNet-100. The heatmap visualizes the pairwise cosine similarity between feature activation vectors across the test set. The Cross-Entropy baseline (left) exhibits diffuse off-diagonal correlations, indicating significant entanglement between feature dimensions. In contrast, CCAR (right) effectively suppresses this interference, yielding a sharp block-diagonal topology that maximizes feature orthogonality.

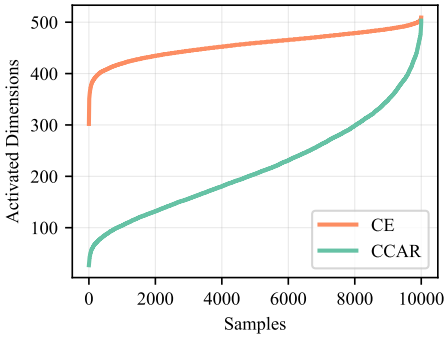


Figure 4. The number of activated dimensions of the representations learned by CE and CCAR on CIFAR-100. CCAR (Green) significantly reduces the effective dimensionality, indicating that the model relies on a compact, sparse set of features for prediction compared to the dense activations of the Cross-Entropy baseline.

$P_{\mathcal{O}_c}$ denotes the orthogonal projection onto subspace $\mathcal{O}_c$, see Appendix B.1), minimizing the expected CCAR-L2 loss is equivalent to minimizing the trace of the class-conditional covariance matrix projected onto the forbidden subspace $\mathcal{O}_c$.

Proof of this property is provided in Appendix B.1.

Remark 4.2. This theorem establishes a direct link to Fisher Discriminant Analysis (FDA) (Fisher, 1936). The Fisher Ratio is defined as $J = \frac{\mathrm{Tr}(\Sigma_b)}{\mathrm{Tr}(\Sigma_w)}$, where Σ_b is the inter-class scatter and Σ_w is the intra-class scatter. By strictly penalizing $\sum_{j \in \mathcal{O}_c} h_j^2$, CCAR forces the diagonal elements of Σ_c (which constitute Σ_w) to zero for all forbidden indices. Crucially, it does not constrain the variance within the active subspace $\mathcal{I}_c$. This selective compression minimizes the denominator of the Fisher Ratio without collapsing the useful intra-class diversity required for fine-grained discrimination. Unlike standard FDA, which seeks a linear projection after feature extraction, CCAR embeds this optimality directly into the encoder's topology.

4.2. Perturbation Stability

We now turn to the robustness implications of this geometry. A key failure mode of deep networks is sensitivity to small, adversarial perturbations that cross decision boundaries (Szegedy et al., 2014; Goodfellow et al., 2014). We show that CCAR creates a geometric margin against such shifts.

To analyze the stability of the latent manifold, we consider the model in its converged state. We assume the linear classifier W has reached a state of alignment with the partitioned subspaces $\mathcal{P}$, such that each weight vector w_k resides in its designated subspace $\mathcal{I}_k$.

Theorem 4.3 (Perturbation Stability Bound)**.** *Ideally, let a clean feature $h \in \mathcal{I}_y$. For an additive perturbation δ to induce misclassification into target class $k \neq y$, the perturbation magnitude projected onto forbidden subspace $\mathcal{I}_k$ must exceed the geometric margin defined by the ratio of correct class signal to target class weight sensitivity.*

See Appendix B.2 for the proof.

Corollary 4.4 (Certified Robustness)**.** *Let the minimum geometric margin be defined as $\tau = \min_{k \neq y} \frac{w_y^\top h}{\|w_y - w_k\|_2}$. If the perturbation norm satisfies $\|\delta\|_2 < \tau$ and clean features are confined to $\mathcal{I}_y$, then the classifier is provably robust to misclassification.*

Proof. Misclassification into class k occurs if $(w_k - w_y)^\top (h + \delta) > 0$. Since $w_k \perp h$ and $h \in \mathcal{I}_y$, this simplifies to $(w_k - w_y)^\top \delta > w_y^\top h$. By Cauchy-Schwarz, we have $\|\delta\|_2 \|w_k - w_y\|_2 \geq (w_k - w_y)^\top \delta$. Thus, a successful attack requires $\|\delta\|_2 > \frac{w_y^\top h}{\|w_k - w_y\|_2} \geq \tau$. Conversely, $\|\delta\|_2 < \tau$ guarantees stability. □

Remark on Theoretical Confinement. The margin in Theorem 4.3 is dynamic (depending on $\delta_{\mathcal{I}_y}$). This bound assumes the adversary does not suppress the active subspace signal; if the adversary degrades $\mathcal{I}_y$, the required energy in $\mathcal{I}_k$ drops. While these bounds assume ideal conditions, CCAR applies them via a soft L_2 inductive bias. This permits minor residual variance, but we show empirically in Section 5.2 that the induced geometric concentration remains functionally profound and highly effective at preserving stability.

While the preceding results establish bounds for arbitrary shifts, we further characterize the resilience of the learned geometry to isotropic (directionless) noise.

Table 1. Comparison of Linear Probing Accuracy (%) on CIFAR-100 and ImageNet-100 under symmetric label noise. CCAR demonstrates superior robustness across all noise levels, maintaining high performance even in extreme noise regimes where baselines degrade. Results are averaged over five seeds. (Best results in bold).

Noise (%)	Cross-Entropy	Center Loss	Contrastive (CL)	NCL	**CCAR (Ours)**
CIFAR-100					
0	74.46 ± 0.24	75.39 ± 0.15	75.65 ± 0.31	74.21 ± 0.12	$\mathbf{75.69 \pm 0.18}$
10	72.90 ± 0.11	74.33 ± 0.29	75.26 ± 0.22	73.68 ± 0.09	$\mathbf{75.41 \pm 0.25}$
20	72.10 ± 0.35	74.19 ± 0.13	75.39 ± 0.16	73.28 ± 0.28	$\mathbf{75.26 \pm 0.07}$
30	71.52 ± 0.18	73.75 ± 0.32	75.23 ± 0.29	73.10 ± 0.15	$\mathbf{75.26 \pm 0.33}$
40	70.97 ± 0.27	74.15 ± 0.08	74.94 ± 0.24	73.17 ± 0.36	$\mathbf{75.25 \pm 0.12}$
50	70.11 ± 0.13	73.58 ± 0.25	74.84 ± 0.10	72.62 ± 0.21	$\mathbf{75.15 \pm 0.29}$
60	68.59 ± 0.42	73.23 ± 0.33	74.71 ± 0.19	72.61 ± 0.14	$\mathbf{75.11 \pm 0.22}$
70	66.95 ± 0.21	73.33 ± 0.17	73.94 ± 0.35	72.37 ± 0.26	$\mathbf{75.00 \pm 0.11}$
80	63.37 ± 0.38	72.21 ± 0.21	73.87 ± 0.13	71.26 ± 0.30	$\mathbf{74.67 \pm 0.27}$
90	52.74 ± 0.25	69.70 ± 0.41	72.34 ± 0.26	67.27 ± 0.19	$\mathbf{74.21 \pm 0.34}$
ImageNet-100					
0	77.64 ± 0.15	78.34 ± 0.22	79.84 ± 0.09	79.14 ± 0.31	$\mathbf{80.62 \pm 0.14}$
10	76.38 ± 0.29	77.66 ± 0.11	79.50 ± 0.25	78.88 ± 0.17	$\mathbf{80.76 \pm 0.21}$
20	76.22 ± 0.13	77.28 ± 0.34	79.52 ± 0.19	78.68 ± 0.22	$\mathbf{80.60 \pm 0.10}$
30	76.08 ± 0.36	77.12 ± 0.18	79.36 ± 0.27	79.02 ± 0.13	$\mathbf{80.54 \pm 0.28}$
40	76.20 ± 0.22	76.86 ± 0.29	79.34 ± 0.14	78.54 ± 0.35	$\mathbf{80.52 \pm 0.08}$
50	75.68 ± 0.19	77.16 ± 0.12	79.56 ± 0.30	78.56 ± 0.26	$\mathbf{80.42 \pm 0.31}$
60	75.42 ± 0.31	76.84 ± 0.24	79.54 ± 0.16	78.64 ± 0.09	$\mathbf{80.42 \pm 0.23}$
70	75.24 ± 0.14	76.92 ± 0.37	79.44 ± 0.21	78.80 ± 0.18	$\mathbf{80.14 \pm 0.12}$
80	73.66 ± 0.26	75.82 ± 0.15	79.28 ± 0.33	78.56 ± 0.24	$\mathbf{80.28 \pm 0.36}$
90	69.08 ± 0.39	74.08 ± 0.28	79.50 ± 0.12	78.26 ± 0.38	$\mathbf{79.98 \pm 0.64}$

Lemma 4.5 (Dimensional Noise Filtering)**.** *Let $\delta \in \mathbb{R}^D$ be a perturbation sampled from $\mathcal{N}(0, \frac{\sigma^2}{D} I)$. For a class-conditional subspace $\mathcal{I}_k$ of dimension $K = D/C$, the projected noise energy $\|\delta_{\mathcal{I}_k}\|^2$ follows a scaled χ^2 distribution. As $D \to \infty$ with C fixed, provided the geometric margin satisfies $\tau^2 > \frac{\sigma^2}{C}$, the probability that the perturbation bridges the margin vanishes exponentially:*

$$P(\|\delta_{\mathcal{I}_k}\|^2 \geq \tau^2) \leq \exp(-D \cdot \mathcal{A}) \tag{7}$$

where $\mathcal{A}$ is a strictly positive rate constant depending on τ^2 and σ^2.

Proof of this concentration property is provided in Appendix B.3.

Geometric Interpretation. In standard training, decision boundaries are often close to the data manifold in arbitrary directions (Fawzi et al., 2018). Theorem 4.3 implies that CCAR essentially concentrates the class manifold within $\mathcal{I}_y$, pushing the decision boundaries for all other classes k far away in the orthogonal directions $\mathcal{I}_k$. To flip a prediction, an adversary cannot simply perturb the input slightly; they must inject significant energy specifically into the subspace $\mathcal{I}_k$. As the dimension D increases, the probability of random noise aligning with a specific forbidden subspace $\mathcal{I}_k$ diminishes, providing inherent robustness to isotropic noise (Fawzi et al., 2016).

5. Empirical Analysis

We evaluate CCAR across three axes: robustness to supervision corruption, geometric mechanism verification, and emergent input stability. Because representation quality is difficult to measure directly, we empirically probe the learned geometry using label noise, input corruptions, and retrieval accuracy. We demonstrate that while standard Cross-Entropy (CE) training ignores latent topology, CCAR imposes a stable manifold structure that persists even under extreme noise regimes. Implementation details are in Appendix G.

5.1. Robustness to Supervision Corruption

Standard empirical risk optimization often causes encoders to memorize stochastic associations when labels are corrupted (Zhang et al., 2017). By treating feature entangle-

ment as a structural vulnerability, we posit CCAR provides a mechanistic defense against such overfitting by confining class-specific information to disjoint orthogonal subspaces. We evaluate this on CIFAR-100 (Krizhevsky et al., 2009) and ImageNet-100 (Deng et al., 2009) under 0% to 90% symmetric label noise.

Baselines. We benchmark against standard Cross-Entropy (CE) (Shannon, 1948) alongside strong geometric and contrastive alternatives: Center Loss (Wen et al., 2016) (minimizes intra-class variance), SimCLR (Chen et al., 2020) (self-supervised contrastive learning), and Non-negative Contrastive Learning (NCL) (Wang et al., 2024) (enforces sparsity). Implementation details are provided in Appendix G.5.

Results. As visualized in Figure 2(a) and summarized in Table 1, CCAR demonstrates superior stability across critical noise regimes. On clean data (0% noise), CCAR achieves comparable accuracy to the baselines, indicating that the subspace constraints ($K = D/C$) do not inhibit representational capacity. However, as the noise rate increases, the divergence becomes pronounced.

On CIFAR-100, Center Loss and Contrastive methods degrade significantly at high noise levels, whereas CCAR maintains robust discriminative power, outperforming baselines by a wide margin at 80% noise. On ImageNet-100, the unregularized baseline collapses, but CCAR effectively filters the high-energy signals required to fit random labels. Crucially, CCAR is complementary to semi-supervised filtering pipelines like DivideMix (Li et al., 2020); while DivideMix requires complex multi-stage training to clean data, CCAR provides a passive, training-time structural regularizer. Scaling to ResNet-50 (Appendix D.4) shows CCAR at 90% corruption outperforms unconstrained baselines trained on pristine data. Finally, CCAR excels on the real-world CIFAR-100N benchmark (Appendix D.3) and, as a proof-of-concept transferability check, yields consistent mAP gains when applied without modification to a YOLOv8 backbone on an aerial object detection task (Appendix D.5).

5.2. Geometric Mechanism Verification

We verify whether the observed robustness arises from the predicted geometric structural changes by analyzing the latent topology through three lenses: semantic consistency, feature sparsity, and subspace partitioning (detailed metrics in Appendix G.2).

Disentanglement and Sparsity. CCAR fundamentally shifts feature organization (Figure 2). Unlike the CE baseline's diffuse, entangled correlations, CCAR enforces a sharp block-diagonal topology (Figure 3). Quantitatively (Figure 2c), CCAR achieves substantially higher sample-level feature sparsity ($\approx 58\%$ on CIFAR-100, $\approx 28\%$ on ImageNet-100) than the baseline ($\approx 11\%$ and $< 2\%$). Crucially, this sparsity is non-destructive: CCAR yields a significantly higher Class Consistency Rate (Figure 2b), implying the active neurons remain highly semantically relevant to the input class.

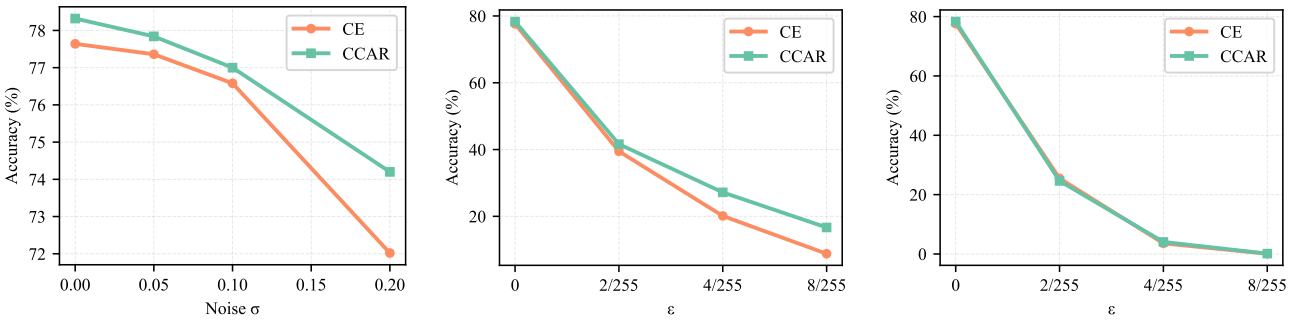
Quantifying Subspace Confinement. To confirm this sparsity aligns with Theorem 4.3, we evaluate how tightly the soft inductive bias confines features. While the baseline utilizes nearly all dimensions per sample, CCAR maintains a compact, block-sparse profile (Figure 4). We rigorously quantify this via the Subspace Energy Density Ratio (Appendix E.2). On ImageNet-100 (ResNet-50), the CE baseline distributes energy uniformly (ratio $\approx 0.9\times$), whereas CCAR concentrates energy within the designated subspace at $194.4\times$ the density of forbidden regions. This profound structural concentration physically bridges our theoretical bounds and practical training.

Functional Purity via Masked Inference. To prove this concentrated energy holds the true semantic signal, we perform an ablated prediction test. Note that Masked Inference uses the ground-truth label at inference (an oracle test) and is strictly a diagnostic probe of geometric structure, not a predictive accuracy claim. Applying the class mask $M(y)$ at inference destroys the entangled CE baseline (accuracy plummets to near-chance). Conversely, CCAR's accuracy improves significantly (detailed in Appendix E.2), confirming the soft constraint successfully captures the core semantic signal within the target subspace, rendering residual leakage functionally irrelevant.

5.3. Emergent Input Stability and Retrieval

Finally, to evaluate if the geometric margin (Theorem 4.3) and statistical filtering (Lemma 4.5) extend to unseen perturbations, we test models initially trained under 40% symmetric label noise (Figure 5) to explicitly isolate emergent stability under stress. We subject these models to three corruptions: isotropic Gaussian noise ($\sigma \in \{0.05, 0.1, 0.2\}$), single-step adversarial perturbations (FGSM) (Goodfellow et al., 2014) at $\epsilon \in \{2/255, 4/255, 8/255\}$, and iterative adversarial attacks (PGD-20) (Madry et al., 2018) with step size $\alpha = \epsilon/4$.

Stability Analysis. Under stochastic Gaussian noise and single-step FGSM attacks (Figures 5a,b), CCAR demonstrates a distinct stability advantage over the CE baseline. This empirically supports Theorem 4.3: minimizing forbidden energy naturally increases the perturbation budget required to cross class boundaries. However, under iterative PGD-20 attacks (Figure 5c), CCAR performs comparably to the baseline. Because CCAR employs a soft inductive bias rather than a rigid architectural projection, minor activation energy inevitably leaks into forbidden dimensions. While wide geometric margins easily absorb single-step and



(a) Isotropic Gaussian Noise (b) Single-Step FGSM Attack (c) Iterative PGD-20 Attack

Figure 5. Robustness analysis on ImageNet-100. We compare the classification accuracy of CCAR against the CE baseline under **(a)** Isotropic Gaussian Noise, **(b)** Single-step FGSM perturbations, and **(c)** Iterative PGD-20 attacks.

stochastic noise, iterative attacks systematically accumulate gradients to navigate this residual leakage.

Semantic Retrieval Quality. Robust learning often trades feature discriminability for stability. We investigate this via semantic image retrieval (mAP@10). CCAR breaks this trade-off, achieving significantly higher mAP scores than the baseline (Figure 6). Here, the x-axis denotes evaluation on truncated feature subsets (retaining only the top-d most active dimensions during inference), proving CCAR concentrates semantic quality into fewer dimensions. Thus, explicitly decoupling class subspaces enhances the features' metric quality. Visualizations are provided in Appendix F.2.

6. Conclusion

Standard supervised learning often yields unstructured, entangled representations. Identifying this geometric entanglement as a critical vulnerability, we propose Class-Conditional Activation Regularization (CCAR) as a targeted mechanistic defense. With minimal computational overhead and no additional data, CCAR-L2 imposes a block-diagonal covariance structure, confining class-specific information to orthogonal subspaces. This significantly enhances robustness against both supervision corruption and unseen input perturbations. Theoretically, we established formal links between forbidden energy minimization and Fisher Discriminant Ratio maximization. Empirically, this geometric organization maintains high discriminative power under extreme label noise regimes where standard Empirical Risk Minimization collapses. Ultimately, explicitly shaping latent topology serves as a powerful inductive bias, demonstrating that intrinsic robustness can be engineered directly into the network's geometric foundation.

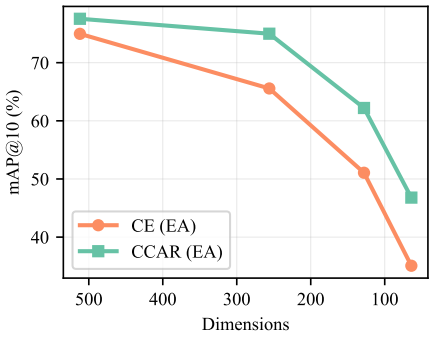


Figure 6. Image retrieval with CE and CCAR features on ImageNet-100. CCAR achieves a significantly higher mAP@10 score, indicating that geometric disentanglement improves semantic consistency alongside robustness.

Limitations. Operationally, CCAR neutralizes stochastic noise and single-step perturbations, but is not a standalone defense against iterative attacks. Furthermore, CCAR requires $D \geq C$; if $D < C$, the subspace dimension $K = \lfloor D/C \rfloor = 0$, preventing out-of-the-box use on setups like ImageNet-1K with a ResNet-18 bottleneck ($D = 512$). This requires grouping classes into superclasses or adding a projection head. Finally, against complex instance-dependent noise, CCAR is best coupled with dedicated data-cleaning pipelines.

Impact Statement. By enhancing model stability without the massive computational overhead of adversarial training, CCAR offers an environmentally sustainable path toward safe and secure AI. The enforced feature disentanglement also aids mechanistic interpretability, facilitating transparent auditing. However, a highly robust model may still confidently deploy biased predictions if the underlying dataset contains systemic demographic flaws. Therefore, geometric regularization must always be paired with rigorous fairness auditing prior to deployment.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. In *International Conference on Learning Representations*, 2016.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6541–6549, 2017.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8):1798–1828, 2013.

Bertsekas, D. P. *Nonlinear programming*. Athena scientific, 1999.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.

Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le, Q. V. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 113–123, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Deng, J., Guo, J., Xue, N., and Zafeiriou, S. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., et al. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Fawzi, A., Moosavi-Dezfooli, S.-M., and Frossard, P. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Fawzi, A., Moosavi-Dezfooli, S.-M., Frossard, P., and Soatto, S. Empirical study of the topology and geometry of deep networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3762–3770, 2018.

Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ghosh, A., Kumar, H., and Sastry, P. S. Robust loss functions under label noise for deep neural networks. In *AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jing, L., Zbontar, J., and LeCun, Y. Implicit rank-minimizing autoencoder. In *NeurIPS*, 2020.

Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., and Krishnan, D. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lezama, J., Qiu, Q., and Sapiro, G. Olé: Orthogonal low-rank embedding for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Li, J., Socher, R., and Hoi, S. C. Dividemix: Learning with noisy labels as semi-supervised learning. In *International Conference on Learning Representations*, 2020.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Nair, V. and Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 807–814, 2010.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 2020.

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

Shannon, C. E. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5265–5274, 2018.

Wang, Y., Zhang, Q., Guo, Y., and Wang, Y. Non-negative contrastive learning. In *International Conference on Learning Representations*, 2024.

Wen, Y., Zhang, K., Li, Z., and Qiao, Y. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pp. 499–515. Springer, 2016.

Zeiler, M. D. and Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833. Springer, 2014.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

A. Extended Related Work

While Section 2 outlines the primary motivations for Class-Conditional Geometric Regularization (CCAR), this appendix provides a more granular discussion of its relationship to adjacent fields, specifically Metric Learning, Neural Collapse, and Disentanglement.

A.1. Relationship to Metric Learning and Margin Losses

Metric learning aims to learn a distance function such that similar samples are close and dissimilar ones are far. Seminal works like Center Loss (Wen et al., 2016) penalize the Euclidean distance between deep features and their class centroids, minimizing intra-class variance. ArcFace (Deng et al., 2019) and CosFace (Wang et al., 2018) enforce angular margins on the hypersphere to maximize inter-class separability.

Contrast with CCAR. While these methods effectively structure the feature space, they operate on relative metrics (Euclidean/Angular) rather than absolute coordinate representation. For example, ArcFace (Deng et al., 2019) allows class clusters to rotate arbitrarily. CCAR fundamentally differs by enforcing a low-rank orthogonal subspace structure. By forcing features into disjoint subspaces, CCAR prevents dimensional entanglement. While axis-alignment is a practical parameterization, any orthogonal rotation of the partition yields equivalent geometric margins, as a subsequent linear classifier losslessly absorbs the rotation. This low-rank constraint enables emergent robustness by explicitly suppressing noise in forbidden dimensions.

A.2. Relationship to Neural Collapse

Recent theoretical work has identified a phenomenon known as Neural Collapse (Papyan et al., 2020), which occurs during the terminal phase of training with Cross-Entropy loss. Neural Collapse is characterized by (i) intra-class variation vanishing (features collapse to the class mean), and (ii) class means forming an optimal Simplex Equiangular Tight Frame (ETF).

Contrast with CCAR. While Neural Collapse describes the natural tendency of unregularized networks to compress features, it typically leads to a Simplex geometry where all dimensions are coupled. CCAR actively intervenes in this process to enforce a Block-Diagonal geometry instead of a Simplex ETF. By explicitly penalizing off-diagonal energy (via Theorem 4.1), CCAR prevents the distribution of information across all dimensions, preserving a structured sparsity that the natural Neural Collapse process does not guarantee. This structured collapse is crucial for robustness, as it leaves margin regions of near-zero energy that absorb adversarial perturbations.

A.3. Relationship to Disentangled Representation Learning

Disentanglement learning, often explored via VAEs (e.g., β-VAE) (Higgins et al., 2017), aims to separate distinct generative factors of variation (e.g., pose, color) into independent latent dimensions.

Contrast with CCAR. Standard disentanglement is typically unsupervised and seeks to separate latent attributes. CCAR can be viewed as a form of Supervised Disentanglement, where the factor of variation being disentangled is the class identity itself. This addresses the theoretical impossibility of unsupervised disentanglement without inductive biases (Locatello et al., 2019). Unlike VAEs, which rely on reconstruction objectives and KL-divergence penalties, CCAR achieves this separation discriminatively through the subspace masking mechanism. This avoids the trade-off between reconstruction quality and disentanglement often seen in generative models, allowing CCAR to maintain high classification accuracy while ensuring that class identity is perfectly disentangled from other nuisance variations in the forbidden subspaces.

A.4. Relationship to Adversarial Training

Adversarial Training (Madry et al., 2018) is the gold standard for robustness, formulating training as a min-max game against a worst-case perturbation.

Contrast with CCAR. Adversarial Training is computationally expensive (requiring iterative gradient steps to generate attacks) and often overfits to the specific perturbation model (e.g., L_∞ ball). CCAR offers an efficient alternative: rather than simulating attacks, it constructs a geometry of resistance. By minimizing the dimensionality of the manifold occupied by each class, CCAR statistically reduces the probability that a random or directed perturbation will align with the target class subspace. While Adversarial Training builds a wall at the specific boundary, CCAR moves the data manifold away from the decision boundaries entirely, providing a more general form of stability at negligible computational cost.

B. Proofs

B.1. Proof of Theorem 4.1

Proof. Let $P_{\mathcal{O}_c} = \text{diag}(M(c))$ be the orthogonal projection matrix onto the forbidden subspace indices for class c. The CCAR-L2 objective for a feature vector h is defined as $\mathcal{L}(h, c) \propto \|P_{\mathcal{O}_c} h\|_2^2$. We expand the expected loss over the class-conditional distribution:

$$\begin{aligned}
\mathbb{E}_{h|c}[\mathcal{L}(h,c)] &= \mathbb{E}_{h|c}\left[(P_{\mathcal{O}_c}h)^\top (P_{\mathcal{O}_c}h)\right] && (8)\\
&= \mathbb{E}_{h|c}\left[h^\top P_{\mathcal{O}_c}^\top P_{\mathcal{O}_c} h\right] && (9)\\
&= \mathbb{E}_{h|c}\left[h^\top P_{\mathcal{O}_c} h\right] \quad \text{(Idempotence: } P^\top P = P) && (10)\\
&= \mathbb{E}_{h|c}\left[\text{Tr}(h^\top P_{\mathcal{O}_c} h)\right] \quad \text{(Scalar trace identity)} && (11)\\
&= \mathbb{E}_{h|c}\left[\text{Tr}(P_{\mathcal{O}_c} h h^\top)\right] \quad \text{(Cyclic property)} && (12)\\
&= \text{Tr}\left(P_{\mathcal{O}_c}\mathbb{E}_{h|c}[hh^\top]\right) \quad \text{(Linearity of Expectation)} && (13)
\end{aligned}$$

We invoke the definition of the second moment matrix $\mathbb{E}[XX^\top] = \text{Cov}(X) + \mathbb{E}[X]\mathbb{E}[X]^\top$. Substituting the class covariance Σ_c and class mean μ_c:

$$\begin{aligned}
\text{Tr}\left(P_{\mathcal{O}_c}(\Sigma_c + \mu_c\mu_c^\top)\right) &= \text{Tr}(P_{\mathcal{O}_c}\Sigma_c) + \text{Tr}(P_{\mathcal{O}_c}\mu_c\mu_c^\top) && (14)\\
&= \text{Tr}(P_{\mathcal{O}_c}\Sigma_c) + \text{Tr}(\mu_c^\top P_{\mathcal{O}_c}\mu_c) && (15)\\
&= \text{Tr}(P_{\mathcal{O}_c}\Sigma_c) + \|P_{\mathcal{O}_c}\mu_c\|_2^2 && (16)
\end{aligned}$$

The objective decouples into two terms: (1) the trace of the projected covariance and (2) the projected energy of the mean.

$$\mathbb{E}[\mathcal{L}] \propto \text{Tr}(P_{\mathcal{O}_c}\Sigma_c) + \|P_{\mathcal{O}_c}\mu_c\|_2^2 \tag{17}$$

Under the alignment assumption stated in Theorem 4.1 (where $\|P_{\mathcal{O}_c}\mu_c\|^2 \to 0$), the second term vanishes. Thus, minimizing the loss becomes equivalent to minimizing the covariance trace:

$$\min \mathbb{E}_{h|c}[\mathcal{L}(h,c)] \iff \min \text{Tr}(P_{\mathcal{O}_c}\Sigma_c). \tag{18}$$

□

B.2. Proof of Theorem 4.3

Proof. Let w_y and w_k be the weight vectors for the correct class y and target class k. We analyze the system at the converged state where subspace alignment holds ($w_c \in \mathcal{I}_c$).

Justification for Alignment: Under the idealized assumption that features are perfectly confined to their active subspace at convergence, the weight alignment follows geometrically. Consequently, optimization ensures w_c aligns with $\mathcal{I}_c$.

Assuming this alignment and ideal feature placement $h \in \mathcal{I}_y$ (thus $w_k^\top h = 0$), a misclassification occurs under perturbation δ if:

$$w_k^\top(h+\delta) > w_y^\top(h+\delta) \tag{19}$$

Expanding the inner products:

$$\begin{aligned}
w_k^\top h + w_k^\top\delta &> w_y^\top h + w_y^\top\delta && (20)\\
0 + w_k^\top\delta &> w_y^\top h + w_y^\top\delta \quad \text{(Orthogonality)} && (21)
\end{aligned}$$

We decompose δ relative to the subspace partition $\mathcal{P}$:

$$\delta = \delta_{\mathcal{I}_k} + \delta_{\mathcal{I}_y} + \sum_{j\neq k,y} \delta_{\mathcal{I}_j} \tag{22}$$

Substituting this decomposition into the weight-perturbation dot products:

$$w_k^\top \delta = w_k^\top \delta_{\mathcal{I}_k} + w_k^\top \delta_{\mathcal{I}_y} + \cdots = w_k^\top \delta_{\mathcal{I}_k} \quad (\text{since } w_k \perp \mathcal{I}_{j\neq k}) \tag{23}$$

$$w_y^\top \delta = w_y^\top \delta_{\mathcal{I}_y} + w_y^\top \delta_{\mathcal{I}_k} + \cdots = w_y^\top \delta_{\mathcal{I}_y} \tag{24}$$

The inequality simplifies to:

$$w_k^\top \delta_{\mathcal{I}_k} > w_y^\top h + w_y^\top \delta_{\mathcal{I}_y} \tag{25}$$

Applying the Cauchy-Schwarz inequality ($|u^\top v| \leq \|u\|\|v\|$) to the LHS term $w_k^\top \delta_{\mathcal{I}_k}$:

$$\|w_k\|\|\delta_{\mathcal{I}_k}\| \geq w_k^\top \delta_{\mathcal{I}_k} > w_y^\top (h + \delta_{\mathcal{I}_y}) \tag{26}$$

Dividing by $\|w_k\|$ yields the lower bound on the forbidden energy:

$$\|\delta_{\mathcal{I}_k}\| > \frac{w_y^\top h + w_y^\top \delta_{\mathcal{I}_y}}{\|w_k\|} \tag{27}$$

This confirms that for misclassification to occur, the projection of noise onto the forbidden subspace $\|\delta_{\mathcal{I}_k}\|$ must exceed the geometric margin defined by the RHS. □

B.3. Proof of Lemma 4.5

Proof. Let the perturbation vector $\delta \in \mathbb{R}^D$ be sampled from an isotropic Gaussian distribution $\delta \sim \mathcal{N}(0, \frac{\sigma^2}{D} I_D)$. We analyze the energy projected onto a specific class-conditional subspace $\mathcal{I}_k$ with cardinality $|\mathcal{I}_k| = K = D/C$. The projection $\delta_{\mathcal{I}_k}$ consists of K independent components $\delta_j \sim \mathcal{N}(0, \frac{\sigma^2}{D})$. Standardizing these variables as $z_j = \frac{\sqrt{D}}{\sigma}\delta_j \sim \mathcal{N}(0, 1)$, we express the projected energy as:

$$\|\delta_{\mathcal{I}_k}\|^2 = \sum_{j\in\mathcal{I}_k} \delta_j^2 = \frac{\sigma^2}{D}\sum_{j=1}^{K} z_j^2 = \frac{\sigma^2}{D} Q_K, \tag{28}$$

where Q_K follows a Chi-squared distribution with K degrees of freedom. We seek to bound the probability that this energy exceeds the geometric margin τ^2:

$$P(\|\delta_{\mathcal{I}_k}\|^2 \geq \tau^2) = P\left(Q_K \geq \frac{D\tau^2}{\sigma^2}\right). \tag{29}$$

Applying the Chernoff bound with parameter $t > 0$:

$$P\left(Q_K \geq \frac{D\tau^2}{\sigma^2}\right) \leq \inf_{0<t<1/2} \frac{\mathbb{E}[e^{tQ_K}]}{e^{t\frac{D\tau^2}{\sigma^2}}} \tag{30}$$

$$= \inf_{0<t<1/2} \exp\left(-t\frac{D\tau^2}{\sigma^2} - \frac{K}{2}\ln(1-2t)\right). \tag{31}$$

Minimizing the exponent $g(t)$ with respect to t yields the optimal parameter t^*:

$$\frac{\partial g}{\partial t} = -\frac{D\tau^2}{\sigma^2} + \frac{K}{1-2t} = 0 \implies t^* = \frac{1}{2}\left(1 - \frac{K\sigma^2}{D\tau^2}\right). \tag{32}$$

Substituting $K = D/C$, we have $t^* = \frac{1}{2}(1 - \frac{\sigma^2}{C\tau^2})$. The condition $\tau^2 > \frac{\sigma^2}{C}$ ensures $0 < t^* < 1/2$. Substituting t^* back into the bound:

$$P(\|\delta_{\mathcal{I}_k}\|^2 \geq \tau^2) \leq \exp\left(-\frac{D\tau^2}{2\sigma^2}\left(1 - \frac{\sigma^2}{C\tau^2}\right) - \frac{D}{2C}\ln\left(\frac{\sigma^2}{C\tau^2}\right)\right) \tag{33}$$

$$= \exp\left(-D\left[\frac{\tau^2}{2\sigma^2} - \frac{1}{2C} - \frac{1}{2C}\ln\left(\frac{C\tau^2}{\sigma^2}\right)\right]\right) \tag{34}$$

$$= \exp\left(-D\left[\frac{1}{2C}\left(\frac{C\tau^2}{\sigma^2} - 1 - \ln\frac{C\tau^2}{\sigma^2}\right)\right]\right). \tag{35}$$

Defining the rate constant $\mathcal{A} = \frac{1}{2C}(\frac{C\tau^2}{\sigma^2} - 1 - \ln\frac{C\tau^2}{\sigma^2})$, which is strictly positive for $\frac{C\tau^2}{\sigma^2} > 1$, we conclude:

$$P(\|\delta_{\mathcal{I}_k}\|^2 \geq \tau^2) \leq \exp(-D \cdot \mathcal{A}). \tag{36}$$

Applying the Chernoff bound with parameter $t > 0$:

$$P\left(Q_K \geq \frac{D\tau}{\sigma^2}\right) \leq \inf_{0<t<1/2} \frac{\mathbb{E}[e^{tQ_K}]}{e^{t\frac{D\tau}{\sigma^2}}} \tag{37}$$

$$= \inf_{0<t<1/2} \exp\left(-t\frac{D\tau}{\sigma^2} - \frac{K}{2}\ln(1-2t)\right). \tag{38}$$

Minimizing the exponent $g(t)$ with respect to t yields the optimal parameter t^*:

$$\frac{\partial g}{\partial t} = -\frac{D\tau}{\sigma^2} + \frac{K}{1-2t} = 0 \implies t^* = \frac{1}{2}\left(1 - \frac{K\sigma^2}{D\tau}\right). \tag{39}$$

Substituting $K = D/C$, we have $t^* = \frac{1}{2}(1 - \frac{\sigma^2}{C\tau})$. The condition $\tau > \frac{\sigma^2}{C}$ ensures $0 < t^* < 1/2$. Substituting t^* back into the bound:

$$P(\|\delta_{\mathcal{I}_k}\|^2 \geq \tau) \leq \exp\left(-\frac{D\tau}{2\sigma^2}\left(1 - \frac{\sigma^2}{C\tau}\right) - \frac{D}{2C}\ln\left(\frac{\sigma^2}{C\tau}\right)\right) \tag{40}$$

$$= \exp\left(-D\left[\frac{\tau}{2\sigma^2} - \frac{1}{2C} + \frac{1}{2C}\ln\left(\frac{C\tau}{\sigma^2}\right)^{-1}\right]\right) \tag{41}$$

$$= \exp\left(-D\left[\frac{1}{2C}\left(\frac{C\tau}{\sigma^2} - 1 - \ln\frac{C\tau}{\sigma^2}\right)\right]\right). \tag{42}$$

Defining the rate constant $\mathcal{A} = \frac{1}{2C}(\frac{C\tau}{\sigma^2} - 1 - \ln\frac{C\tau}{\sigma^2})$, which is strictly positive for $\frac{C\tau}{\sigma^2} > 1$, we conclude:

$$P(\|\delta_{\mathcal{I}_k}\|^2 \geq \tau) \leq \exp(-D \cdot \mathcal{A}). \tag{43}$$

□

C. Regularization Design Analysis

In Section 3.2, we motivated the selection of the squared L_2 norm for the geometric penalty based on its optimization stability and smooth gradient field. To justify this design, we provide a comparative analysis of several candidate regularizers representing distinct geometric inductive biases. We evaluate these variants based on their ability to maintain standard accuracy while providing stability under severe supervision corruption.

C.1. Experimental Configuration

For the supervision corruption tasks, we first train the ResNet-18 (He et al., 2016) backbone from scratch for 200 epochs on CIFAR-100 and ImageNet-100 (Deng et al., 2009). For the experiments on CIFAR-100, we employ a batch size of 256, a weight decay of 0.0005, and an initial learning rate of 0.1 regulated by a Cosine Annealing schedule. For ImageNet-100, we utilize a batch size of 256 and a weight decay of 0.0001 while maintaining the same learning rate and scheduler settings. We apply the regularization penalty $\lambda = 3.0$ from the initial epoch. As noted in the main text, the value of the regularization coefficient λ was determined through an extensive grid search over $\lambda \in \{0.1, 0.5, 1.0, 2.0, 3.0, 4.0, 5.0\}$. While all tested values strictly outperformed the unregularized baseline—confirming the method is not brittle to hyperparameter selection—we fixed $\lambda = 3.0$ as it provided the optimal balance between preserving standard classification performance and maximizing robustness against the 90% high-intensity label noise. To ensure absolutely fair comparison, all baseline methods were configured using the optimal hyperparameter settings reported in their respective original publications (e.g., center weight $\lambda = 0.003$ for Center Loss, temperature $\tau = 0.07$ for SimCLR/NCL), ensuring they operate at their established state-of-the-art capacity. Linear evaluation is performed by freezing the backbone and training a linear classifier for 50 epochs across a noise spectrum ranging from 0% to 90% in 10% increments.

C.2. Dynamics of Geometric Penalty Variants

The L_1 **Sparsity Baseline** serves as a standard choice for inducing sparsity within the representation (Tibshirani, 1996). While it effectively isolates feature activations, it is non-differentiable at the origin, which can potentially destabilize the optimization process during the refinement of fine-grained features.

$$\mathcal{L}_{L_1} = \|M(y) \odot h\|_1 \tag{44}$$

The **Cosine Margin Penalty** borrows principles from margin-based metric learning. This formulation penalizes representations based on the cosine similarity between the feature vector and competing class centroids relative to the target centroid, incorporating a margin δ to enforce separation. The class centroids μ_k are initialized randomly and updated as exponential moving averages (EMA) of the normalized features across training batches. We apply a clamped hinge loss with margin $\delta = 0.35$:

$$\mathcal{L}_{\text{Margin}} = \max\left(0, \max_{k \neq y}(\cos(h, \mu_k)) - \cos(h, \mu_y) + \delta\right) \tag{45}$$

The **Subspace Energy Ratio** constrains the logarithmic ratio of energy residing in the forbidden subspace relative to the energy within the active subspace. This approach focuses on the relative distribution of energy across the feature manifold rather than penalizing absolute activation magnitudes.

$$\mathcal{L}_{\text{Ratio}} = \log\left(1 + \frac{\|h_{\text{off}}\|_2^2}{\|h_{\text{aligned}}\|_2^2 + \epsilon}\right) \tag{46}$$

where $h_{\text{aligned}} = (\mathbf{1} - M(y)) \odot h$ is the feature energy inside the active subspace, $h_{\text{off}} = M(y) \odot h$ is the energy in the forbidden region, and $\epsilon = 1e^{-8}$ is added for numerical stability.

The **Orthogonal Projection Penalty** represents a direct subspace constraint. It penalizes any component of the feature vector h that resides in the orthogonal complement of the class-specific subspace, as defined by a projection matrix P_y derived from class-specific statistics. Here, $P_y = U_y U_y^\top$ is computed per-class using the top-K eigenvectors (U_y) of the empirical covariance matrix, which is updated iteratively per batch using a momentum of 0.9.

$$\mathcal{L}_{\text{Proj}} = \|(I - P_y)h\|_2^2 \tag{47}$$

The **Squared L_2 Penalty (Selected Design)** utilizes a smooth quadratic penalty to suppress activations in forbidden dimensions. By indexing these dimensions via a class-conditional mask $M(y)$, the objective provides a restorative force that aligns the latent topology with the semantic hierarchy of the dataset.

$$\mathcal{L}_{L_2} = \frac{1}{|\mathcal{O}_y|} \sum_{j \in \mathcal{O}_y} h_j^2 \tag{48}$$

Results. As shown in Table 2, the L_2 penalty consistently outperforms all alternative geometric constraints across the noise spectrum. For example, at 90% noise, the L_2 formulation achieves 74.21% accuracy, significantly surpassing the L_1 baseline (69.14%) and the runner-up Subspace Energy Ratio (73.29%). This indicates that while sparsity-inducing (L_1) or relative (L_{Ratio}) penalties offer some resilience, the smooth gradient field of the quadratic penalty provides superior optimization stability under extreme corruption. Therefore, we conclude that the L_2 norm is the optimal design choice for implementing the CCAR objective, as it maximizes robustness without compromising the representational capacity required for clean classification.

C.3. Partition Design Ablation: Contiguous vs. Random Assignment

In Section 3.1, CCAR assigns class-specific dimensions using a contiguous layout for practical convenience. Theorem 4.3 relies solely on the disjoint nature of the active subspaces, implying that the specific coordinate ordering is arbitrary. To empirically verify this, we perform an ablation on CIFAR-100 comparing our standard contiguous assignment against a random-fixed assignment (where the K dimensions for each class are selected uniformly at random without replacement).

As shown in Table 3, the random assignment performs within the variance of the contiguous assignment across all noise levels. This confirms that the robustness gains of CCAR are fundamentally driven by the induced orthogonal subspace structure, rather than any artifact of the contiguous coordinate layout.

Table 2. Comparison of Linear Probing Accuracy (%) across geometric penalty variants on CIFAR-100. By transposing the view, we observe the stability of each method as noise increases. The L_2 Penalty consistently outperforms sparsity (L_1) and metric-based constraints. Results are averaged over five seeds.

NOISE (%)	L_1 BASELINE	COSINE MARGIN	ENERGY RATIO	ORTHOGONAL PROJ.	L_2 **PENALTY**
0	73.08±0.21	74.73±0.39	74.13±0.32	75.65±0.28	**75.69±0.18**
10	72.99±0.15	74.62±0.12	74.10±0.36	75.02±0.28	**75.41±0.25**
20	72.88±0.11	74.46±0.39	73.96±0.35	74.99±0.16	**75.26±0.07**
30	72.80±0.16	74.51±0.19	73.92±0.26	75.07±0.23	**75.26±0.33**
40	72.93±0.28	74.46±0.14	74.03±0.19	75.08±0.21	**75.25±0.12**
50	72.78±0.34	74.28±0.16	73.85±0.25	74.97±0.28	**75.15±0.29**
60	72.73±0.28	74.25±0.15	73.74±0.12	74.73±0.38	**75.11±0.22**
70	72.62±0.34	74.06±0.19	73.81±0.13	74.69±0.31	**75.00±0.11**
80	72.36±0.14	73.81±0.25	73.73±0.11	74.02±0.37	**74.67±0.27**
90	69.14±0.30	72.75±0.19	73.29±0.26	72.89±0.26	**74.21±0.34**

Table 3. Linear Probing Accuracy (%) on CIFAR-100 comparing Contiguous vs. Random subspace assignments. Results are averaged over 5 seeds.

ASSIGNMENT	0% NOISE	20% NOISE	40% NOISE	60% NOISE	80% NOISE	90% NOISE
Contiguous	75.69 ± 0.18	75.26 ± 0.07	75.25 ± 0.12	75.11 ± 0.22	74.67 ± 0.27	74.21 ± 0.34
Random	75.71 ± 0.23	75.14 ± 0.19	75.31 ± 0.16	74.98 ± 0.27	74.73 ± 0.31	74.08 ± 0.38

D. Additional Empirical Analysis

In Section 4, we established that minimizing the CCAR objective theoretically compresses the covariance trace in forbidden directions (Theorem 4.1) and partitions the latent space into orthogonal subspaces (Theorem 4.3). In this section, we provide empirical verification of these geometric phenomena using diagnostic metrics computed on the ImageNet-100 validation set.

D.1. Fisher Discriminant Analysis

To quantitatively monitor the emergence of the block-diagonal structure, we track the Fisher Discriminant Ratio (FDR) throughout the training trajectory. The FDR serves as a proxy for linear separability, defined as the ratio of the inter-class scatter trace to the intra-class scatter trace (Fisher, 1936): $J = \text{Tr}(\Sigma_b)/\text{Tr}(\Sigma_w)$.

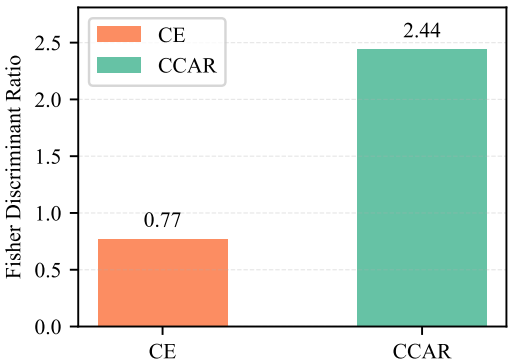


Figure 7. Fisher Ratio analysis on CIFAR-100. CCAR ($J = 2.44$) shows significantly stronger class separation than CE ($J = 0.77$).

Results. As illustrated in Figure 7, CCAR maintains a significantly higher Fisher Ratio compared to the Cross-Entropy (CE) baseline throughout the training process. Notably, while the FDR for the baseline degrades rapidly as supervision noise increases, CCAR exhibits robust separability. This confirms that the explicit energy minimization penalty continuously suppresses intra-class variance (Σ_w), effectively maximizing the FDR even when the label signal is unreliable.

D.2. Verification of Covariance Compression

Theorem 4.1 states that the objective minimizes the trace of the intra-class covariance matrix in forbidden directions. We verify the effectiveness of this minimization by analyzing the Mean Cluster Radii for the 100 classes in ImageNet-100.

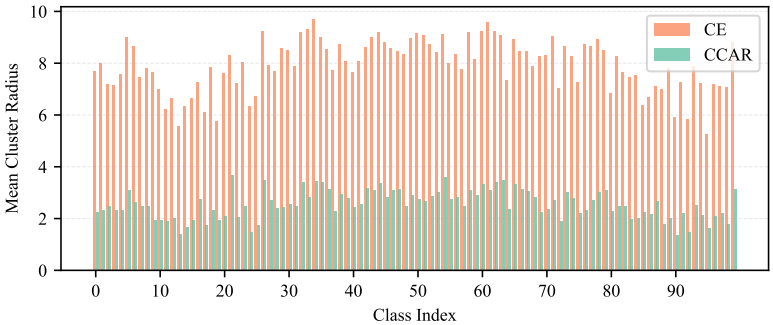


Figure 8. Impact of CCAR on Intra-Class Compactness: Mean Cluster Radii Analysis on ImageNet-100. CCAR yields consistently lower radii compared to the baseline, confirming that the regularization effectively compresses intra-class variance and induces tighter class manifolds.

Results. As illustrated in Figure 8, the clusters formed by CCAR are generally more compact than those of the baseline. The reduction in mean radius indicates that the regularization successfully compresses intra-class variance. This compression appears to selectively target the directions of variation that do not align with the class-specific subspaces, resulting in tighter class manifolds that are spatially distant from the decision boundaries of competing classes.

D.3. Robustness to Real-World Asymmetric Noise (CIFAR-100N)

In Section 5.1, we evaluated CCAR using symmetric label noise to cleanly isolate and validate the geometric mechanism. However, real-world label corruption is often asymmetric and instance-dependent. To verify that the geometric margins induced by CCAR provide robustness in real-world scenarios, we evaluate the method on the CIFAR-100N dataset, which contains real human-annotated noisy labels.

To ensure a rigorous comparison of intrinsic representation quality, we utilized the exact same ResNet-18 backbone and training protocol (200 epoch pretraining, 50 epoch linear probe) as the symmetric noise experiments. Critically, we applied the same $\lambda = 3.0$ regularization strength without any dataset-specific hyperparameter tuning.

As shown in Table 4, CCAR outperforms the Cross-Entropy baseline on real-world noise. More importantly, the variance across 5 random seeds is drastically reduced (from ± 0.28 to ± 0.08). This indicates that the orthogonal subspace constraints force the network into a highly stable geometric configuration that consistently resists overfitting to complex, human-generated noise patterns.

Table 4. Linear Probing Accuracy (%) on the real-world CIFAR-100N benchmark. Results are averaged over 5 seeds.

METHOD	CIFAR-100N ACCURACY (%)
Control (CE)	52.75 ± 0.28
CCAR (Ours)	$\mathbf{53.81 \pm 0.08}$

D.4. Architecture Scaling: ResNet-50 on ImageNet-100

To ensure that the geometric benefits of CCAR are not specific to a single architecture, we evaluated the method's scalability by employing a higher-capacity ResNet-50 backbone. On ImageNet-100, the expanded feature dimension ($D = 2048$) increases the per-class subspace dimension K from 5 to 20, representing a 4x increase in available capacity per class compared to the ResNet-18 setting.

As detailed in Table D.4, CCAR effectively leverages this expanded capacity. The performance gap between CCAR and the Cross-Entropy baseline grows monotonically with the noise level, confirming that the geometric confinement mechanism

scales cleanly to larger architectures without modification.

Most notably, the CCAR model trained under the extreme 90% label noise regime (79.98%) achieves a higher linear probing accuracy than the standard Cross-Entropy baseline trained on entirely clean, uncorrupted data (79.36%). This demonstrates a profound property of latent geometry: structural topology can supersede data purity. By explicitly enforcing an orthogonal scaffold, the network distills a purer semantic signal from severe noise than standard, unconstrained optimization extracts from ideal conditions.

Table 5. Linear Probing Accuracy (%) using a ResNet-50 backbone on ImageNet-100. The performance gap grows monotonically, and CCAR at 90% noise remarkably outperforms the CE baseline at 0% noise. Results are averaged over five seeds.

NOISE (%)	CONTROL (CE)	**CCAR (OURS)**	GAP
0	79.36 ± 0.21	80.64 ± 0.18	+1.28
10	78.68 ± 0.25	80.82 ± 0.15	+2.14
30	78.44 ± 0.28	80.54 ± 0.19	+2.10
50	78.20 ± 0.35	80.42 ± 0.22	+2.22
70	77.30 ± 0.42	80.14 ± 0.25	+2.84
80	76.62 ± 0.51	80.28 ± 0.28	+3.66
90	74.16 ± 0.73	79.98 ± 0.31	+5.82

D.5. Downstream Transfer: Dense Prediction with YOLOv8

To establish CCAR as a general-purpose, plug-and-play geometric regularizer, we must ensure that the induced orthogonal subspace structure does not destroy the spatial feature maps required for dense prediction tasks. To verify downstream transferability, we integrated the CCAR penalty into a YOLOv8 backbone and evaluated it on a 5-class aerial object detection task on a custom internal real-world dataset.

Crucially, all hyperparameters were left entirely at their YOLOv8 defaults. CCAR was added solely as an additional penalty term on the backbone features, with no modifications made to the detection head, anchor assignments, or training schedule.

As shown in Table 6, the geometric regularization of the backbone yielded significant downstream improvements across all detection metrics. Most notably, we observe an +8.27% improvement in mAP@50-95 and an +8.11% boost in Precision. This confirms that the disentangled representations learned via CCAR seamlessly transfer to complex localization tasks, proving the method's viability as a foundational enhancement for modern vision pipelines.

Table 6. Object Detection performance using a YOLOv8 backbone on a real-world aerial dataset. Integrating CCAR yields significant improvements across all metrics without any task-specific tuning.

METRIC	YOLOV8 BASELINE	**YOLOV8 + CCAR**	GAP
mAP@50	93.40%	97.14%	+3.74%
mAP@50-95	59.41%	67.68%	+8.27%
Precision	87.59%	95.70%	+8.11%
Recall	88.87%	93.78%	+4.91%

Scope and intent. This experiment is designed as a **proof-of-concept transferability check**, not a reproducibility benchmark. Its sole purpose is to test whether the geometric regularization validated on CIFAR-100 and ImageNet-100 produces measurable performance gains when applied without modification to a structurally different task (multi-class object detection on aerial imagery). Because the dataset is an internal collection, we do not claim reproducibility on this specific split; the result demonstrates *generality of effect* rather than state-of-the-art detection performance. A fully reproducible detection benchmark on a named public dataset (e.g., VisDrone2019) is a natural and planned extension of this work.

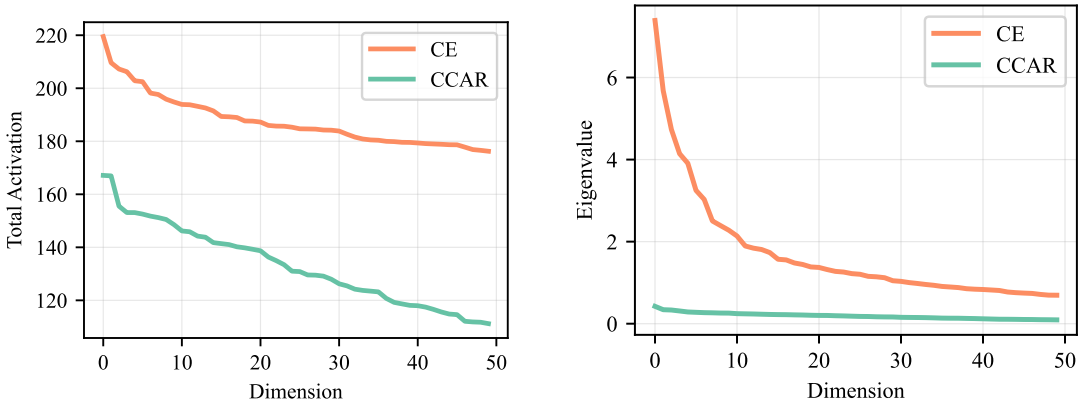
E. Empirical Verifications of Theoretical Analysis on CCAR

In Section 4, we established that minimizing the CCAR objective theoretically compresses the covariance trace in forbidden directions (Theorem 4.1) and partitions the latent space into optimal orthogonal subspaces. In this section, we provide

empirical verification of these phenomena using diagnostic metrics on ImageNet-100.

E.1. Verification of Optimal Representations

To verify the theoretical analysis of the optimal solutions derived in Theorem 4.1 (Trace Minimization), we investigate the internal topology of the learned representations. We examine two complementary metrics: the Expected Activation (EA) distribution, which measures where feature energy is concentrated, and the Eigenvalue Distribution, which measures the effective rank and capacity of the representation.



(a) Distribution of the 50 largest EA values

(b) Distribution of the 50 largest eigenvalues

Figure 9. Analysis of learned features of CE and CCAR on ImageNet-100.

Energy Concentration (EA). As illustrated in Figure 9(a), we observe a distinct difference in energy allocation. The features learned by CCAR exhibit a concentrated activation profile where the majority of signal energy is contained within the top-ranked dimensions. This alignment is consistent with the subspace constraints imposed by our objective, confirming that the network effectively utilizes the designated active dimensions while suppressing variance in the forbidden directions. In contrast, the baseline CE model shows a uniform distribution of energy across the spectrum, indicating a dispersed representation without explicit subspace localization.

Trace Minimization (Eigenvalues). Furthermore, the eigenvalue distribution in Figure 9(b) reveals the spectral signature of trace minimization. The CCAR eigenspectrum exhibits a sharp decay followed by a flat tail near zero, confirming that variance in forbidden directions has been effectively eliminated. Crucially, the spectrum remains flat in the leading dimensions, unlike the baseline which exhibits a slower decay. This suggests that CCAR avoids dimensional collapse (Jing et al., 2020); instead of collapsing to a single point, the representation utilizes the full capacity of the allowed block (I_c) to encode intra-class variations, approximating the ideal block-diagonal geometry described in Theorem 4.1 more closely than the unregularized baseline (Papyan et al., 2020).

E.2. Quantifying Functional Confinement: Diagnostic Methodology

In Section 5.2, we introduced two diagnostic metrics to empirically validate the "near-perfect confinement" assumption of Theorem 4.3 under the soft L_2 inductive bias. The methodologies for computing these metrics on the ImageNet-100 validation set are detailed below.

1. Subspace Energy Density Ratio (DR). Because CCAR confines class representations to highly specific coordinate blocks, measuring raw energy leakage can be misleading due to the massive difference in volume between the assigned subspace (1% of total dimensions) and the forbidden region (99% of total dimensions). To provide a scale-invariant measure of geometric concentration, we calculate the Subspace Energy Density Ratio (DR) for each sample. This is defined as the mean squared activation (energy) per dimension inside the active subspace $\mathcal{I}_y$, divided by the mean squared activation per dimension in the forbidden region $\mathcal{O}_y$:

$$\mathrm{DR}(x, y) = \frac{\frac{1}{|\mathcal{I}_y|} \sum_{j \in \mathcal{I}_y} h_j^2}{\frac{1}{|\mathcal{O}_y|} \sum_{j \in \mathcal{O}_y} h_j^2 + \epsilon} \tag{49}$$

where $\epsilon = 1e^{-12}$ is a numerical stability term. A density ratio near $1.0\times$ (as observed in the CE baseline) indicates that feature energy is uniformly distributed across the entire network capacity. The $194.4\times$ mean ratio observed for CCAR mathematically confirms that the regularizer successfully concentrates activation energy into the designated structural block, closely approximating the theoretical ideal.

2. Masked Inference (Ablated Prediction Test). To ensure that the residual energy located in the forbidden region $\mathcal{O}_y$ does not dictate the network's predictive behavior, we conducted an ablated inference test. During the standard forward pass of the validation set, we intervene immediately before the final fully connected layer (classifier W) by applying a binary mask.

For the *Designated-Only* condition, we apply the ideal class mask $M(y)$ derived from the ground-truth label, explicitly zeroing out all activations in the forbidden dimensions while leaving the K dimensions in $\mathcal{I}_y$ untouched. The fact that the Cross-Entropy baseline drops to near-chance (3.66%) confirms that standard training distributes critical semantic information diffusely across all dimensions. Conversely, the CCAR model not only retains its performance but improves to 88.90%. This confirms that the geometric penalty forces the network to encode the entirety of the necessary discriminative signal exclusively within the designated K dimensions, rendering any residual leakage functionally obsolete.

F. Extended Feature Properties

Beyond the theoretical verification, we investigate the practical properties of the learned features that emerge from this geometric structuring, specifically focusing on intrinsic linearity and semantic interpretability.

F.1. Intrinsic Linearity and Separability

A key theoretical requirement for disentangled representations is intrinsic linear separability: semantic information should be accessible via a single affine transformation, without requiring complex non-linear decoding (Alain & Bengio, 2016). To verify this, we compare the classification accuracy of a Linear Probe against a Multi-Layer Perceptron (MLP) Probe (two layers with ReLU (Nair & Hinton, 2010)) trained on frozen features.

Table 7. Classification accuracy (%) of Linear vs. MLP probes using a ResNet-18 backbone on the ImageNet-100 dataset. The bottom row (Δ) highlights the significant stability gain of CCAR, particularly in the high-noise regime. (Note: Baseline clean accuracy variations from Table 1 are due to probe-specific hyperparameter tuning differences required for a fair MLP comparison).

	Clean Data		**90% Label Noise**	
Method	Linear	MLP	Linear	MLP
Control (CE)	77.38	77.34	61.28	66.36
CCAR (Ours)	**78.16**	**78.20**	**73.18**	**76.20**
Δ	*+0.78*	*+0.86*	*+11.90*	*+9.84*

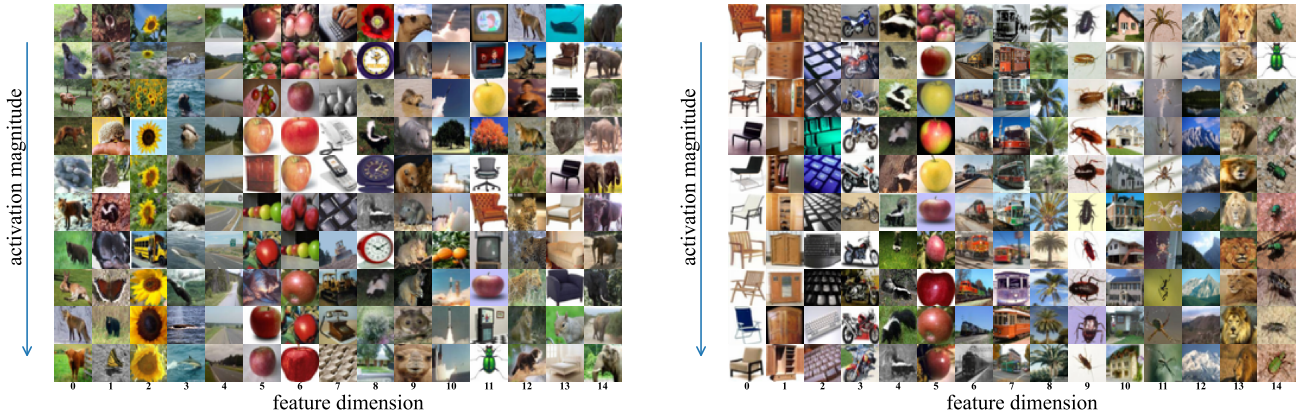
As summarized in Table 7, CCAR consistently outperforms the Cross-Entropy (CE) baseline. On clean data, CCAR achieves a Linear accuracy of 78.16%, surpassing the baseline's 77.38%. This advantage becomes critical in high-noise regimes (90% label noise), where the baseline's linear separability collapses to 61.28%. Under these conditions, the baseline relies heavily on the MLP to recover performance (+5.08% boost), indicating that its feature manifold becomes entangled and curved. In contrast, CCAR maintains a high linear accuracy of 73.18%, with the linear probe capturing the vast majority of the semantic structure. This confirms that the CCAR manifold remains intrinsically flat and robust, even when the supervision signal is heavily corrupted.

F.2. Visual Interpretability

Finally, we provide a qualitative assessment of the semantic information encoded within the learned features. We visualize the top-activating images for specific neurons to determine if the geometric constraints translate into semantic consistency (Zeiler & Fergus, 2014). We select neurons based on their maximum activation values on the test set and display the images that elicit the strongest response, a method akin to Network Dissection (Bau et al., 2017).

The visualizations in Figure 10 and Figure 11 reveal a high degree of semantic coherence in the CCAR representations. We observe that neurons within the assigned subspaces tend to fire exclusively for concepts related to that class. For example, specific neurons within a vehicle-assigned subspace consistently activate on wheel patterns or metallic textures, while

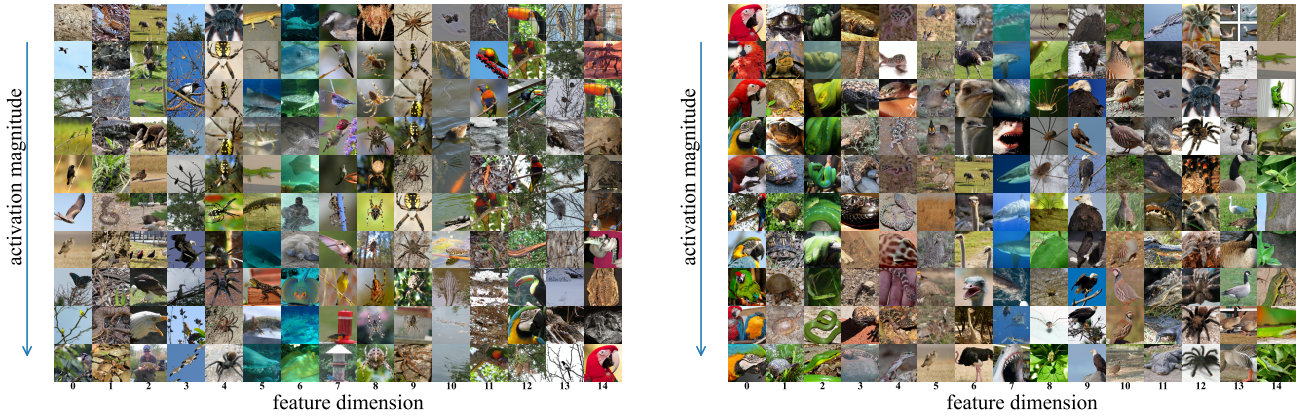
remaining inactive for biologically textured images. This aligns with findings that standard CNNs are biased towards texture rather than shape (Geirhos et al., 2019).



(a) Cross Entropy (b) Class-Conditional Activation Regularization (Ours)

Figure 10. Visualization of CIFAR-100 test samples with the largest values along each feature dimension (sorted according to activation values).

Crucially, this visualization highlights the property of intra-class polysemanticity discussed in Section 3. While the subspace is class-specific, the individual neurons within that subspace capture diverse attributes of the class. One neuron may specialize in the frontal view of an object, while another captures the side profile. This contrasts with the baseline, where top activating images for a single neuron often span multiple semantically unrelated classes. The semantic purity observed in CCAR confirms that the geometric regularization forces the network to align its internal feature axes with meaningful semantic factors, effectively filtering out spurious correlations that often bridge different classes in unregularized models.



(a) Cross Entropy (b) Class-Conditional Activation Regularization (Ours)

Figure 11. Visualization of ImageNet-100 test samples with the largest values along each feature dimension (sorted according to activation values).

G. Details of Experiment Setup

G.1. Implementation and Evaluation Protocol

In our implementation, the encoder f_θ consists of a ResNet-18 backbone g (He et al., 2016). For all experiments, we apply the Class-Conditional Geometric Regularization at the penultimate feature layer $h = g(x)$. The regularization term $\mathcal{L}_{\text{CCAR}}$ is computed by applying class-conditional binary masks $M(y)$ to these activations. We select the regularization strength $\lambda = 3$ based on a pilot study on CIFAR-100 where it demonstrated the optimal balance between subspace compression and gradient stability.

Evaluation. When evaluating representation quality, we discard the final classification weights and utilize the backbone outputs. For linear probing, we train a single linear layer for 50 epochs with a learning rate of 0.1 and Cosine Annealing (Loshchilov & Hutter, 2017). For the manifold linearity experiments in Appendix F, we utilize a two-layer MLP probe with a hidden dimension of 2048 to assess if non-linear decoding can recover information from entangled baseline features.

G.2. Geometric Analysis Implementation

In this section, we define the geometric metrics utilized for the mechanism verification in Section 5.2 and the theoretical proofs in Appendix E. All metrics are computed on the ImageNet-100 validation set using the ResNet-18 backbone trained for 200 epochs.

To visualize the internal topology of the latent manifold, we compute the **Feature Correlation Matrix** (C). Each entry C_{ij} represents the cosine similarity between the activations of feature dimensions i and j, computed over the expectation of the validation set:

$$C_{ij} = \frac{|\mathbb{E}[h_i^\top h_j]|}{\|h_i\|\|h_j\|} \tag{50}$$

To reveal the block-diagonal structure, we sort the feature dimensions according to their assigned class indices. Dimensions with zero activation variance (dead neurons) are excluded from this visualization to ensure the heatmap accurately reflects the geometry of the active manifold.

We quantify the representational efficiency of the encoder using **Feature Sparsity**, which measures the fraction of the feature space that remains inactive for a given input. A dimension h_j is considered inactive if its magnitude falls below a numerical threshold $\epsilon = 1e^{-5}$. The sparsity score $S(x)$ for a sample x is defined as:

$$S(x) = \frac{1}{D}\sum_{j=1}^{D} \mathbb{I}(|h_j(x)| < \epsilon) \tag{51}$$

We report the mean sparsity averaged over the entire test set.

To assess the semantic purity of the learned features, we utilize the **Class Consistency Rate (CCR)**. This metric evaluates whether individual neurons specialize in concepts relevant to their assigned class subspace. For each neuron j, we retrieve the set of top-K maximally activating images $\{x^{(1)}, \dots, x^{(K)}\}$ from the validation set (with $K = 10$). The consistency score is calculated as the proportion of these images that belong to the class c assigned to that neuron's subspace $\mathcal{I}_c$:

$$\text{CCR}_j = \frac{1}{K}\sum_{k=1}^{K} \mathbb{I}(y(x^{(k)}) = c) \tag{52}$$

The global CCR is reported as the mean score across all active neurons.

Finally, to quantitatively monitor the compression of class manifolds referenced in Appendix D, we track the **Fisher Discriminant Ratio (FDR)**. This metric serves as a proxy for linear separability and is defined as the ratio of the trace of the inter-class scatter matrix (Σ_b) to the intra-class scatter matrix (Σ_w) (Fisher, 1936):

$$\text{FDR} = \frac{\text{Tr}(\Sigma_b)}{\text{Tr}(\Sigma_w)} \tag{53}$$

We track this metric at regular intervals during training; high values indicate that the class manifolds have collapsed into compact, widely separated clusters, while low values indicate diffuse and overlapping distributions.

G.3. Experiment Details of Input Stability

To evaluate the perturbation stability properties derived in Theorem 4.3, we subject models trained under 40% symmetric label noise to diverse input-level corruptions. We characterize robustness against random pixel-level distortions by adding isotropic Gaussian noise $\delta \sim \mathcal{N}(0, \sigma^2 I)$ to the normalized test samples. Adversarial stability is assessed through both single-step and iterative optimization-based attacks. We utilize the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2014) with perturbation magnitudes $\epsilon \in \{2/255, 4/255, 8/255\}$.

Iterative robustness is measured using Projected Gradient Descent (PGD) (Madry et al., 2018) with 20 iterations, employing a total budget of $\epsilon = 8/255$ and a step size of $\alpha = 2/255$. All robustness evaluations are performed strictly on the ResNet-18 backbones preserved from the training phase to ensure that the reported metrics reflect the intrinsic stability of the learned geometric manifold rather than the capacity of a specific classification head. It is worth noting that we maintain the same data normalization and resolution parameters used during the training phase to ensure that the results are not biased by variations in the input pipeline.

G.4. Experiment Details of Supervision Corruption

For the supervision corruption tasks, during the training process, we utilize ResNet-18 (He et al., 2016) as the backbone and train the models from scratch for 200 epochs on CIFAR-100 (Krizhevsky et al., 2009), and ImageNet-100 (Deng et al., 2009). For the experiments on CIFAR-100, we employ a batch size of 256, a weight decay of 0.0005, and an initial learning rate of 0.1 regulated by a Cosine Annealing schedule (Loshchilov & Hutter, 2017). For ImageNet-100, we utilize a batch size of 256 and a weight decay of 0.0001 while maintaining the same learning rate and scheduler settings. We apply the regularization penalty $\lambda = 3$ from the initial epoch without any warmup period.

To evaluate robustness, symmetric label noise is injected from 0% to 90% in 10% increments. During the linear evaluation stage, we freeze the backbone and train a classifier for 50 epochs following the standard protocol. For the retrieval experiments, we utilize cosine similarity to identify the top-10 nearest neighbors in the feature space and report the mean Average Precision at 10. It is worth noting that we apply these settings consistently across both datasets to ensure a fair comparison of the geometric properties learned by the model.

G.5. Baseline Implementation Details

To ensure reproducibility and fair comparison within the supervised benchmark, we explicitly define the implementation of the contrastive baselines referenced in Section 5.1. Since SimCLR (Chen et al., 2020) and NCL (Wang et al., 2024) are typically utilized in self-supervised or distinct learning paradigms, we adapt them to the supervised setting as **auxiliary regularization objectives**. This ensures a fair comparison with CCAR, where the geometric constraints are applied alongside the standard supervision. The total loss is formulated as $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{base}}$, where both objectives are optimized jointly from scratch.

- **SimCLR (Contrastive):** We attach a non-linear projection head (2-layer MLP) to the backbone and minimize the NT-Xent loss on the projected features. This objective serves as a regularizer to encourage instance-level discrimination alongside the class-level supervision provided by Cross-Entropy.
- **NCL (Non-negative):** We follow the official implementation (Wang et al., 2024), applying non-negative constraints to the projection head weights and utilizing the NCL objective function to encourage sparsity in the learned representations.
- **Center Loss:** We maintain a memory bank of learnable class centroids. The model minimizes the intra-class Euclidean distance between features and their corresponding class centers, effectively reducing the intra-class variance Σ_w.

Crucially, these methods are not used for pre-training; they are integrated directly into the supervised training loop. All baselines utilize the same ResNet-18 backbone, data augmentation pipeline, and training schedule (200 epochs) as the proposed CCAR method to ensure that differences in performance are attributable to the geometric regularization strategy rather than the training protocol.

G.6. Compute Resources

For all core experiments, including the training of the ResNet-18 and ResNet-50 backbones, models are trained on a local workstation equipped with an Intel Core i7 processor, 64 GB of RAM, and a single NVIDIA GeForce RTX 4080 GPU. Regardless of the regularization method, all models follow the same training configurations and batch sizes. Training a ResNet-18 model from scratch for 200 epochs requires approximately 5.5 hours for the CIFAR-100 dataset and 25 hours for the ImageNet-100 dataset. For the geometric diagnostic tests and ablated inference evaluations detailed in Section 5.2, feature extraction and metric computation are performed on the same hardware, requiring less than 10 minutes per validation set.