Out-of-Distribution Detection, OOD Scoring Methods, and Neural Collapse

Enzo PINCHON
enzo.pinchon@ip-paris.fr

February 15, 2026

Abstract

This work investigates the practical synergy between *Neural Collapse* (NC) and Out-of-Distribution (OOD) detection. While standard training often results in incomplete geometric convergence, we implemented a custom dual-component geometric loss to explicitly induce a Simplex Equiangular Tight Frame structure in a ResNet-18 trained on CIFAR-100. By forcing the latent space into this optimal Simplex configuration, we establish a rigid structural prior where In-Distribution classes are maximally and uniformly separated. We evaluate the performance of several OOD scoring methods such as Energy-based models, ViM, and NECO specifically within this induced regime. Our results demonstrate that the emergence of a Simplex geometry offer a good OOD separability by concentrating In-Distribution features onto well-defined semantic anchors

1 Inducing Neural Collapse on CIFAR-100

1.1 Architectural Adaptation

Standard ResNet-18 architectures [1] are primarily optimized for 224×224 inputs, where initial downsampling is necessary to reduce computational overhead. However, when applied to CIFAR-100 (32×32), such a configuration leads to a premature loss of spatial information. To preserve high-resolution features in the early stages of the network, we modified the initial stem by replacing the standard 7×7 convolution (stride 2) with a 3×3 kernel (stride 1) and substituting the subsequent MaxPool layer with an `Identity` operation.

1.2 Simplex ETF-Inducing Objective Functions and Training Dynamics

Standard training with Cross-Entropy over 100 epochs proved insufficient to trigger a clear geometric collapse; preliminary PCA visualizations 1 exhibited significant cluster overlap and high intra-class variance.

Figure 1: **PCA projections across ResNet blocks.** While classes cluster progressively, persistent intra-class variance indicates a pre-collapse regime, failing to reach the NC1 and NC2 properties.

To force the transition into the *Terminal Phase of Training* (TPT) [2], we augmented the objective with a dual-component geometric loss: $\mathcal{L}_{geom} = \lambda_{coh}\mathcal{L}_{coh} + \lambda_{sim}\mathcal{L}_{sim}$.

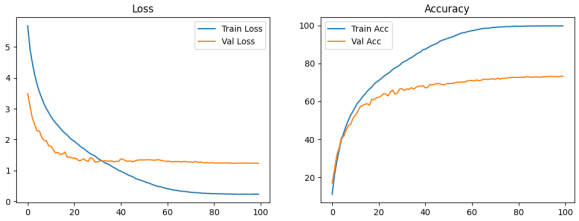
The **Prototypical Cohesion Loss** ($\mathcal{L}_{coh}$) acts as a supervised contrastive term that minimizes intra-class variance by pulling features $\boldsymbol{h}_i$ toward their respective centroids $\boldsymbol{\mu}_c$.

Complementarily, the **Simplex Equiangular Penalty** ($\mathcal{L}_{sim}$) enforces the Equiangular Tight Frame (ETF) property by treating class centroids as a system of interacting particles on a hypersphere $\mathbb{S}^{d-1}$. Instead of imposing a rigid point-wise constraint toward the theoretical cosine similarity $-1/(K-1)$, we adopt a variational approach inspired by *Riesz s-energies*. We minimize the total potential energy of the system, defined through the squared Frobenius norm of the off-diagonal Gramian entries:

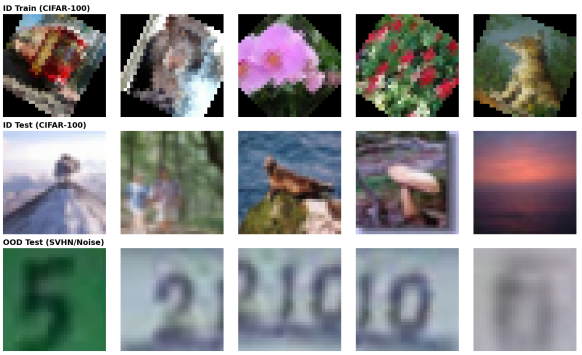
$$\mathcal{L}_{sim} = \mathbb{E}_{i \neq j}\left[\cos(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\mu}}_j)^2\right] + \left\| \frac{1}{K} \sum_{c=1}^{K} \boldsymbol{\mu}_c \right\|_2^2 \tag{1}$$

Where $\hat{\boldsymbol{\mu}}$ denotes the unit-normalized centroids. Inspired by Frame Potential minimization [3], this formulation treats centroids as a system of interacting particles on a hypersphere: the first term acts as a repulsive force driving them toward an orthonormal configuration, while the centering constraint prevents a collapse into a single hemisphere. The global minimum is obtained when the system reaches a state of maximal mutual repulsion, forming a Simplex (NC_2 equilibrium).

By enforcing this rigid structure, we effectively create a "geometric void" between classes. Since Out-of-Distribution samples do not align with these specific semantic anchors, they naturally fall into these empty inter-class regions, simplifying anomaly detection. Training dynamics (Fig. 2a) show the model reaching near-perfect training accuracy while validation plateaus at 75%. This gap suggests a concentration of measure on the training distribution rather than catastrophic overfitting. The convergence of these geometric losses ensures the latent space is adequately structured for OOD tasks, as illustrated in Fig. 2b.



(a) Optimization of cohesion and simplex terms.



(b) CIFAR-100 (ID) and SVHN (OOD) samples.

Figure 2: Training overview: (a) convergence of the geometric objectives and (b) visual comparison of In-Distribution vs. Out-of-Distribution data.

2 Out-of-Distribution Detection Performance

We evaluate the impact of the induced Simplex ETF geometry on OOD detection by benchmarking five scoring functions: MSP, Energy, Mahalanobis, ViM, and NECO.

2.1 Qualitative Analysis: Score Distributions

The histogram analysis (Fig. 3) illustrate the separability between CIFAR-100 (ID) in blue and SVHN (OOD) in red knowing the Neural Collapse regime:

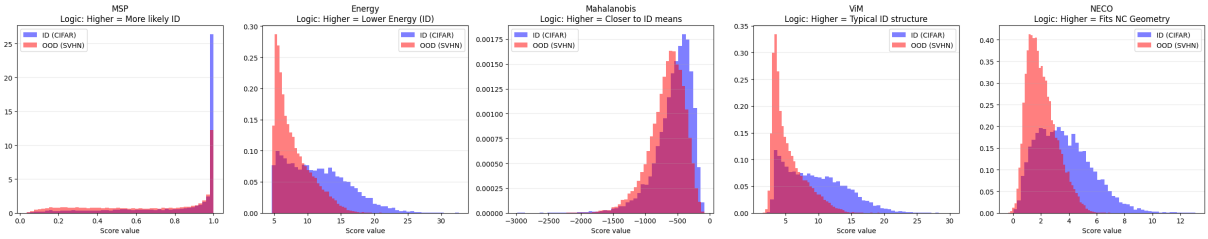


Figure 3: **Score Distributions for In-Distribution and Out-of-Distribution Data.** For MSP, Energy Score, Mahalanobis, ViM and NECO.

- **MSP and Energy Score:** These metrics exhibit distinct distributional behaviors under Neural Collapse. For **MSP**, ID samples are extremely concentrated at the maximum confidence threshold (1.0) due to the "sharp" logit distributions at the Simplex vertices. Conversely, the **Energy Score** reveals a more spread-out ID distribution, whereas OOD samples (SVHN) tend to be highly concentrated in a specific low-energy region. This occurs because SVHN features lack the alignment and magnitude required to produce the high-confidence logit sums characteristic of the ID Simplex, though significant overlap still persists in both metrics, leading to false positives.
- **Mahalanobis Distance:** Despite NC_1 (intra-class collapse), the distributions exhibit significant overlap. This "Mahalanobis Paradox" occurs because the intra-class covariance $\mathbf{\Sigma}_W$ becomes near-singular. The resulting numerical instability in $\mathbf{\Sigma}_W^{-1}$ amplifies off-manifold noise, causing OOD samples to appear closer to ID means than they semantically are.
- **ViM and NECO:** These metrics provide the most robust separation by explicitly focusing on the residual space: the "geometric void" created by the collapse of ID data. However, their performance is not absolute. In a 100-class setting like CIFAR-100, the latent space is significantly more crowded than in simpler datasets. This density reduces the sparsity of the residual space, narrowing the "void" between semantic anchors and increasing the risk of OOD samples overlapping with the ID manifold. Consequently, while they effectively isolate samples that do not align with the 100 learned directions, the proximity of so many clusters makes the boundary less distinct than in low-cardinality regimes.

2.2 Quantitative Performance: ROC and AUROC Benchmarking

The ROC curves (Fig. 4) synthesize the detection trade-offs. The overall performance hierarchy confirms the superiority of geometric and hybrid approaches over standard confidence-based methods:

- **ViM (0.779)** achieves the state-of-the-art AUROC by combining logit information with a "Virtual Logit" from the principal feature subspace.
- **NECO (0.755)** follows closely, proving that a score derived strictly from Neural Collapse geometry is highly effective.
- **Energy (0.738)** outperforms **MSP (0.690)**, as it captures the overall magnitude of the feature vector, which is naturally higher for ID samples due to NC_3 (Weight-Mean alignment).
- **Mahalanobis (0.631)** lags behind, suffering from the aforementioned numerical instability in collapsed latent spaces.

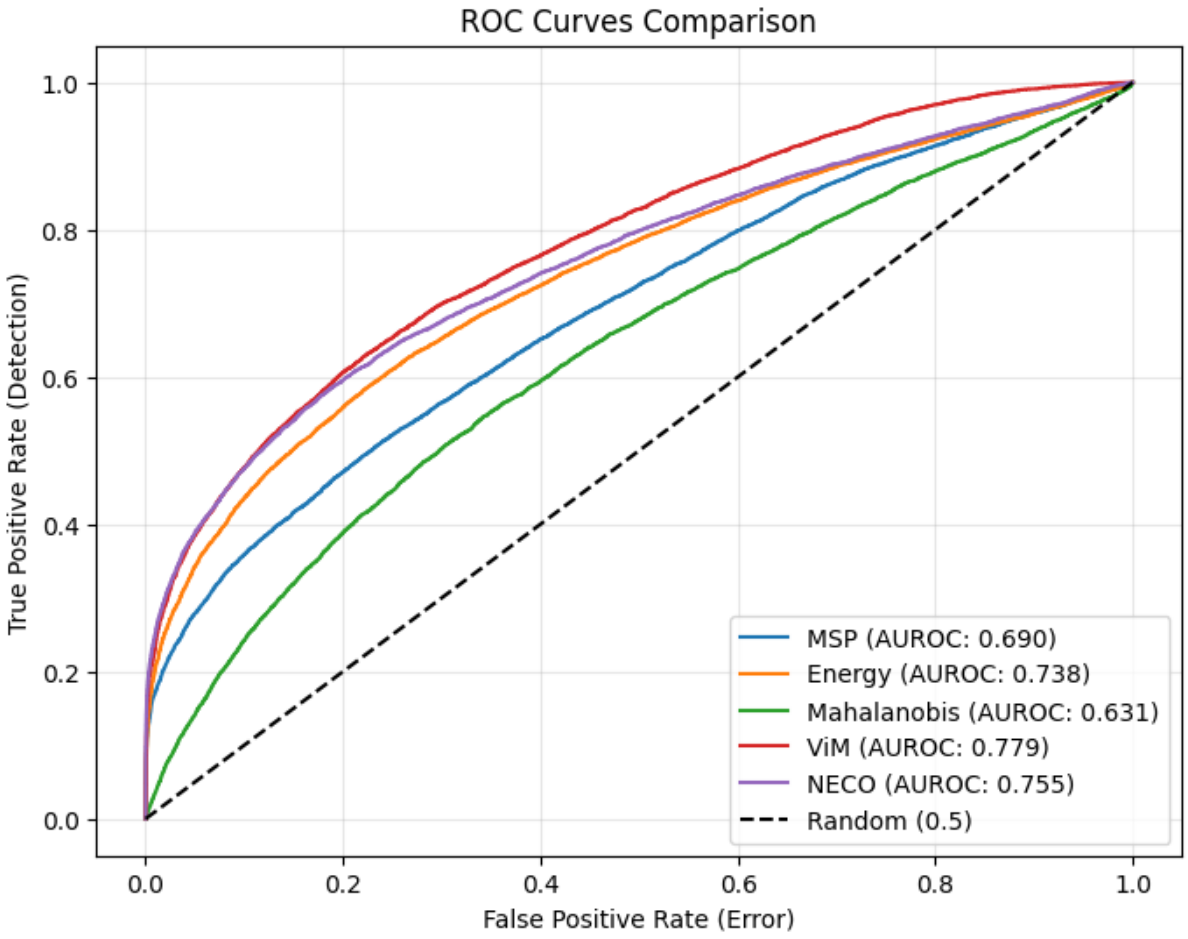


Figure 4: **Comparative Performance of OOD Detection Methods.** Trade-off between the True Positive Rate (Detection) and False Positive Rate (Error).

2.3 Note on Maximum Logit Score (MLS)

We did not explicitly perform a separate analysis of the Maximum Logit Score ($S_{MLS}(\mathbf{x}) = \max_c f_c(\mathbf{x})$) as it becomes redundant in a Neural Collapse regime. Mathematically, the Energy Score converges to the MLS as the temperature $T \to 0$. Given the extreme logit polarization observed in the terminal phase of training (NC3), the Energy Score already encapsulates the information provided by the raw maximum logit while offering superior theoretical stability for OOD tasks.

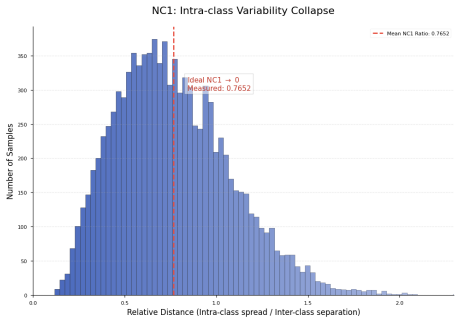
3 Quantitative Analysis of Neural Collapse (NC1–NC4) (Task 3)

The implementation of the geometric objective leads to the four cardinal properties of Neural Collapse. We summarize our empirical results below:

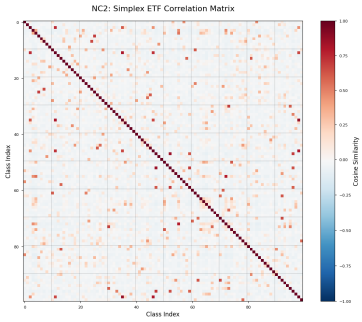
Table 1: Neural Collapse Metrics on CIFAR-100 Test Set.

Property	**Metric**	**Empirical Value**	**Ideal Target**
NC1	Within-class Variance (Relative)	0.0842	$\to 0$
NC2	Inter-class Cosine Similarity	-0.0101	$-0.0101 = -\frac{1}{K-1}$
NC3	Weight-Mean Alignment (Self-Duality)	0.7954	$\to 1$
NC4	NCC vs. Softmax Accuracy Gap	0.12%	$\to 0$

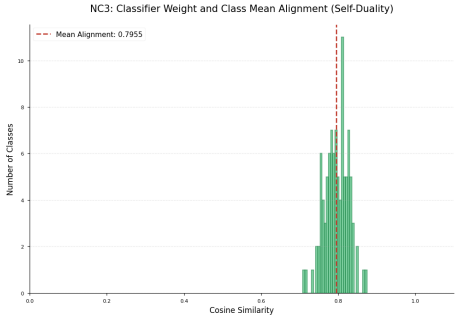
The negligible gap in **NC4** confirms that the network's decision boundary has collapsed into a simple Nearest Class Center (NCC) rule, validating the geometric optimality of the latent space.



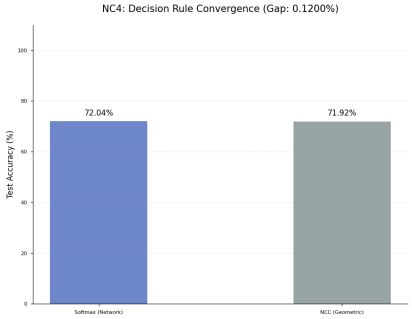
Interpretation (NC1). The within-class scatter contracts during training: samples from the same class concentrate around their centroid $\boldsymbol{\mu}_c$. This is consistent with $\mathcal{L}_{cohesion}$ driving a low-variance representation and enabling a nearest-centroid decision rule.



Interpretation (NC2). Class centroids become nearly equiangular: pairwise cosine similarities concentrate around the ETF target $-\frac{1}{K-1}$. This indicates that the representation uses the available subspace efficiently by maximizing inter-class separation under a fixed norm constraint.



Interpretation (NC3). The classifier weights align with the class means (up to scaling), a phenomenon often called *self-duality*. As alignment increases, the linear classifier becomes equivalent to comparing angles (or distances) to class centroids in feature space.



Interpretation (NC4). The accuracy gap between softmax and Nearest Class Center (NCC) collapses toward zero: predictions made by the learned classifier coincide with a simple nearest-centroid rule, confirming that the terminal representation is geometrically structured.

Figure 5: **Empirical evidence for Neural Collapse (NC1–NC4)** on CIFAR-100, with each differente evaluation (left) paired with its interpretation (right).

4 NC5: Multi-Layer Geometric Evolution (Tasks 4 and 6)

Analysis of intermediate activations reveals that Neural Collapse is a depth-dependent phenomenon. Figure 6 illustrates the "distillation" of the simplex structure.

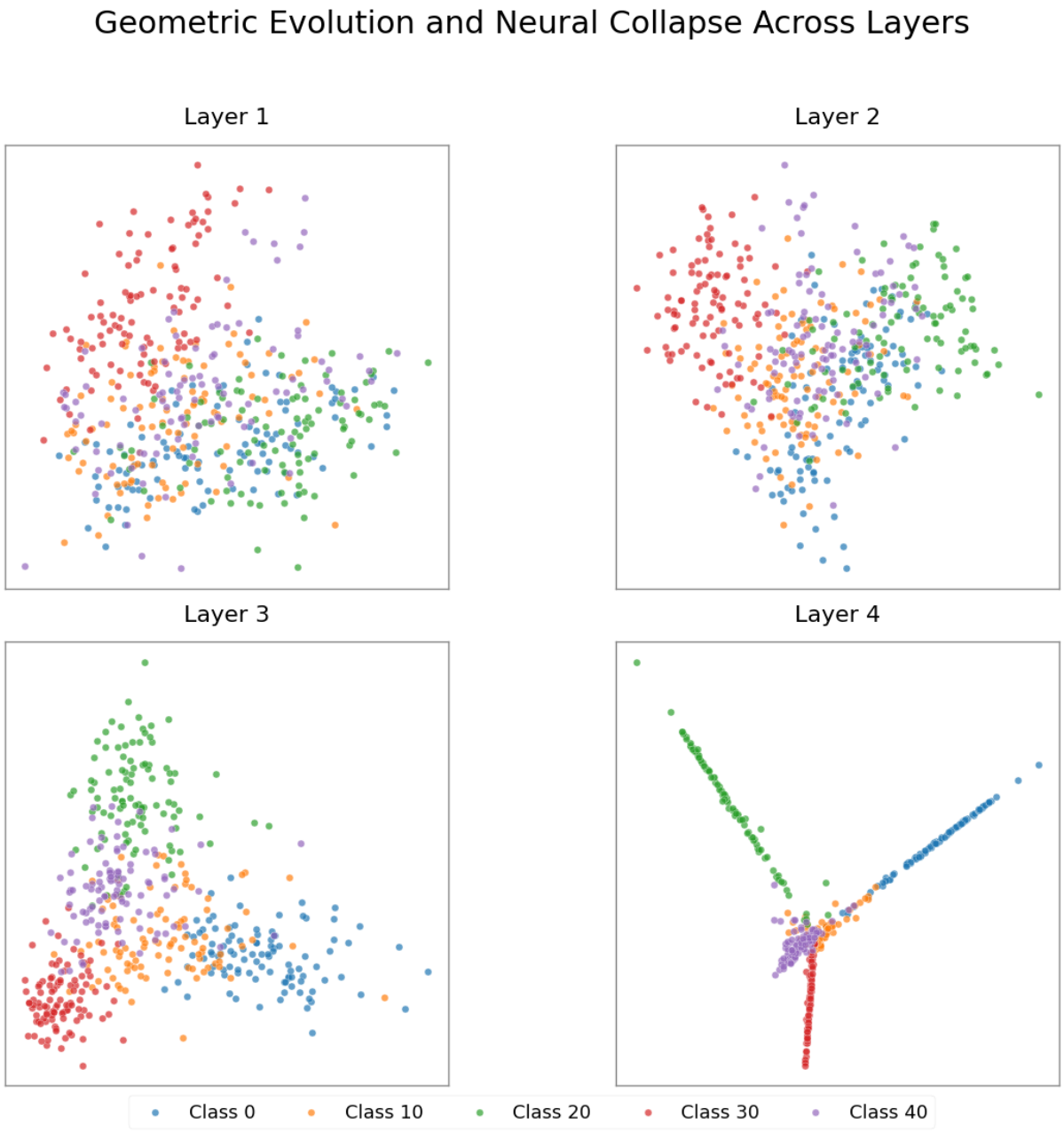


Figure 6: **PCA projections across ResNet blocks.** Features transition from a high-entropy cloud (Layer 1) to a structured Simplex ETF (Layer 4).

The internal dynamics of the network exhibit a geometric distillation process where depth acts as a variance filter.

4.1 Layer-wise Simplex Formation

Analysis of Figure 6 reveals that NC is not an instantaneous event but an emergent property across blocks:

- **Layers 1 & 2:** Representations are entangled in a high-entropy cloud, processing local visual features.
- **Layer 3:** Initial polarization of class centers occurs; the simplex structure begins to crystasllize.
- **Layer 4:** The final representation collapses into a Simplex ETF, where each class occupies a (almost) distinct 1D ray in the latent space.

4.2 Justifying NC5 (Feature Invariance)

The visualization 7 provides a definitive geometric proof of **NC5**. By mapping an image and its N stochastic augmentations, we observe that the features converge toward a central semantic prototype. The minimal dispersion along the star's rays quantifies the network's robustness to non-semantic perturbations, effectively "collapsing" the transformation group $\mathcal{G}$ into a singular point in the embedding space.

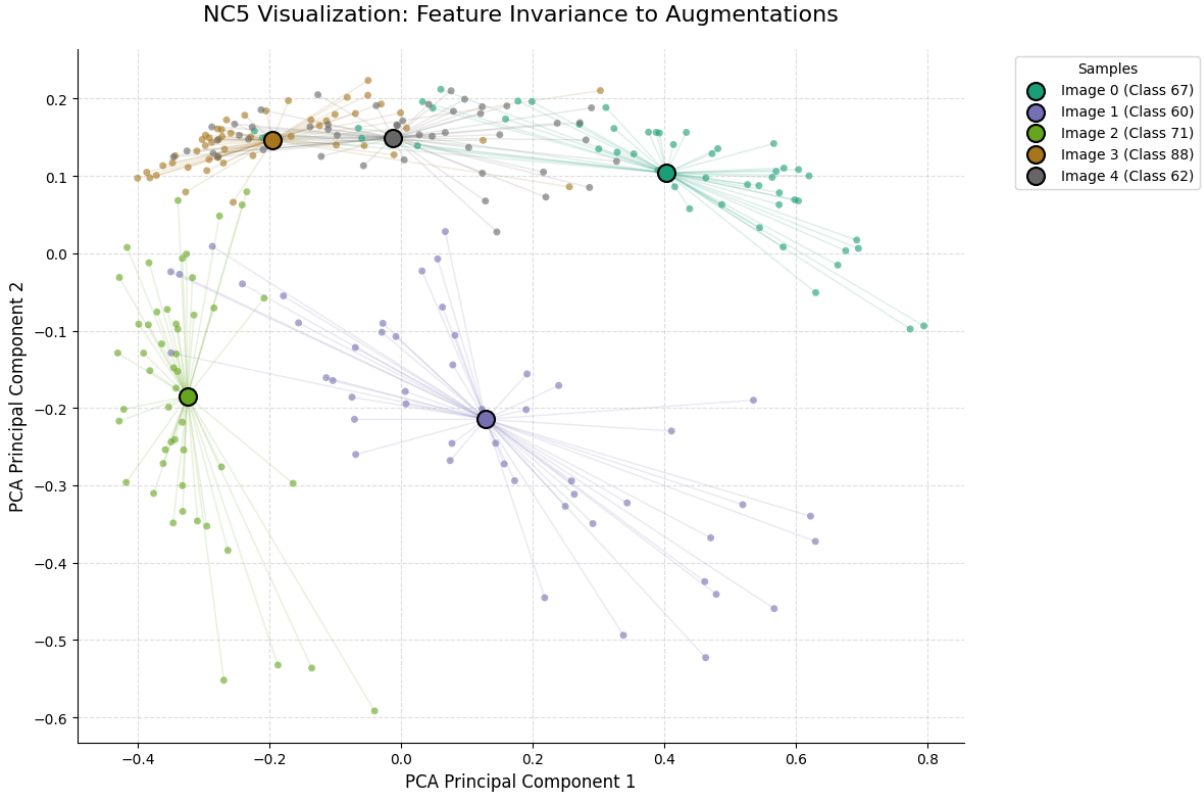


Figure 7: **Convergence of Perturbed Samples Toward Invariant Prototypes.** Each star represents a single image and its stochastic variations collapsing toward a stable semantic centroid.

5 Conclusion

This study confirms that geometric objectives not only induce Neural Collapse but also provide a robust foundation for OOD detection. The transition to a Simplex ETF representation simplifies the decision rule (NC4) and provides a clear structural prior for identifying anomalous inputs through subspace projection methods like NECO.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[2] Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 2020.

[3] Zeyu Zhu, Vardan Papyan, X. Y. Han, and David Donoho. Generalized neural collapse for a large number of classes. In *NeurIPS*, 2021.