Is Complex Training Necessary for Long-Tailed OOD Detection? A Re-think from Feature Geometry

Ningkang Peng **Xuanming Chen** **Yanhui Gu***
Nanjing Normal University
nkpeng@nnu.edu.cn, 19240118@njnu.edu.cn, gu@njnu.edu.cn

Abstract

Long-tailed out-of-distribution (LT-OOD) detection is often addressed with specialized training, including auxiliary out-of-distribution (OOD) data, abstention heads, contrastive objectives, energy losses, or gradient-conflict control. We show that these training mechanisms can obscure a simpler issue: frozen long-tailed representations may already contain useful OOD evidence, but raw Mahalanobis distance is distorted by frequency-coupled feature radius and poorly supported tail covariance. We propose *Hyperspherical Pooled Mahalanobis* (HPM), a post-hoc detector that normalizes features onto the unit sphere and replaces class-specific covariance with a pooled, ridge-regularized metric while keeping class means as semantic anchors. In CIFAR-LT experiments and an ImageNet-100-LT near-OOD evaluation, HPM improves raw class-specific Mahalanobis scoring; for Prior-Calibrated ERM (PC-ERM), it raises AUROC from 45.39 to 85.67 on CIFAR-10-LT and from 54.81 to 78.35 on CIFAR-100-LT. This simple PC-ERM+HPM pipeline also achieves the best Log Efficiency Score (LES; 3.08) in the main CIFAR-100-LT comparison, retaining roughly 95% of that comparison's highest AUROC at substantially lower training-time cost. These results show that simple training can offer a strong performance–cost tradeoff on CIFAR-LT, while ImageNet-100-LT retains a representation-level advantage for specialized training. This motivates separate evaluation of representation quality, detector geometry, and training complexity.

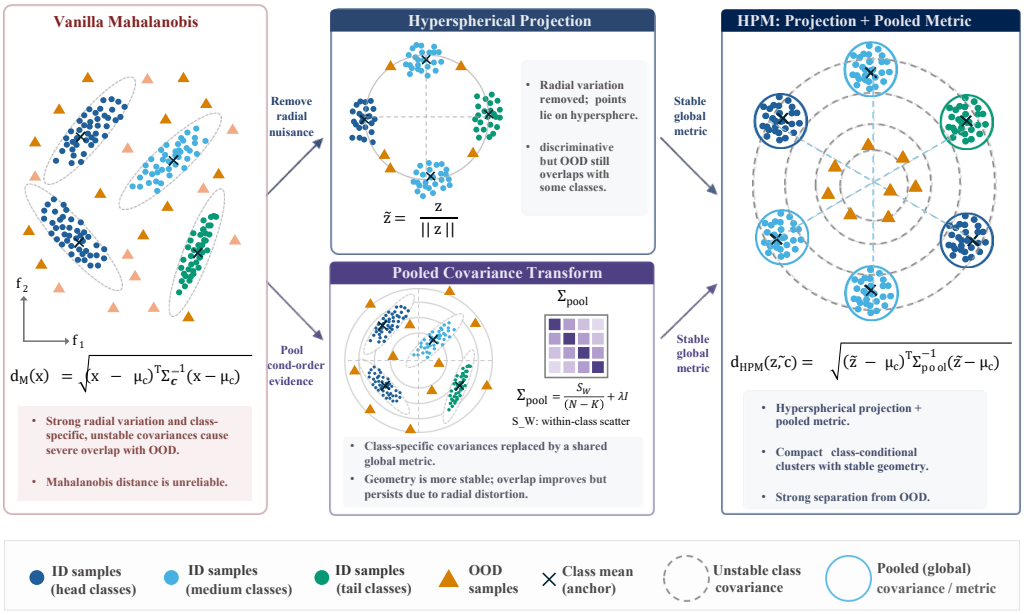
1 Introduction

Open-world recognition must do more than name the classes seen during training. A deployed system also needs to recognize when an input falls outside the in-distribution (ID) training distribution [Hendrycks and Gimpel, 2017, Liang et al., 2018, Lee et al., 2018]. This requirement is especially sharp in long-tailed domains such as autonomous perception, medical triage, ecological monitoring, and industrial inspection, where rare ID categories and novel hazards coexist. Long-tailed out-of-distribution (LT-OOD) detection therefore asks for two behaviors at once: retain rare known classes and reject inputs from outside the training distribution.

Because head classes provide dense supervision while tail classes may have only a few examples [Liu et al., 2019, Cao et al., 2019], many recent LT-OOD methods respond by making training more specialized. Some use energy-based objectives [Liu et al., 2020] or auxiliary outlier exposure [Hendrycks et al., 2019]; others introduce contrastive tail-OOD separation [Wang et al., 2022b], abstention or calibrated outlier classes [Wei et al., 2024, Miao et al., 2024a], adaptive outlier distributions [Miao et al., 2024b], gradient-conflict mitigation [Zhang et al., 2025], or semantic tail prioritization [He et al., 2025]. These methods have advanced the benchmark, but they raise a basic question: is increasingly complex training necessary for strong LT-OOD detection, or can a simple long-tail classifier already be competitive when the detector uses its representation properly?

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Mechanism of Hyperspherical Pooled Mahalanobis

Figure 1: HPM **geometry repair.** Class-conditional Mahalanobis is unstable when tail classes lack residual support (top path). HPM first projects features to the hypersphere to remove radial bias correlated with imbalance, then uses pooled residual covariance to replace tail-specific spectra with shared, ridge-regularized geometry while preserving class mean anchors.

We study this question through Prior-Calibrated ERM (PC-ERM), a simple baseline trained with cross-entropy and long-tail prior calibration [Menon et al., 2021]. PC-ERM uses no auxiliary OOD data, abstention head, contrastive objective, or specialized rejection loss. To compare accuracy with cost, we also report a Log Efficiency Score (LES), which relates the strongest post-hoc OOD AUROC of a trained model to its average per-epoch training time. LES separates absolute OOD performance from compute-normalized performance.

Our experiments point to a different bottleneck: raw Mahalanobis detection reads long-tailed features through an unstable geometry. Feature radius can be coupled with class frequency, so raw distance mixes semantic displacement with a training-frequency artifact. Tail classes also have too few samples to support class-specific covariance, making the estimated precision low-rank, ill-conditioned, and sensitive to regularization. Thus a simple classifier may look weak under raw distance not because its representation is unusable, but because the post-hoc metric is distorted.

We propose *Hyperspherical Pooled Mahalanobis* (HPM), a minimal post-hoc detector that repairs this geometry. It maps features onto the unit hypersphere to remove frequency-coupled radius, then replaces unsupported class-specific covariance with a pooled ridge covariance. Class means remain semantic anchors, while second-order geometry is estimated from shared within-class residuals. Figure 1 summarizes the mechanism. HPM changes only the detector: it adds no outlier head, auxiliary OOD supervision, model retraining, or specialized training objective.

HPM also complements classifier-score diagnostics. Energy and maximum softmax probability (MSP) depend only on classifier outputs, so they cannot respond to feature variation that leaves logits unchanged. This classifier-null blind spot is not our main claim, but it explains why post-hoc LT-OOD evaluation should not rely on classifier scores alone. A repaired feature-space metric can expose evidence that raw Mahalanobis distorts and classifier scores cannot observe.

Our evaluations on CIFAR-10-LT, CIFAR-100-LT, and ImageNet-100-LT use representations from recent LT-OOD and long-tail methods including PATT [He et al., 2025], PASCL [Wang et al., 2022b], AdaptOD [Miao et al., 2024b], COCL [Miao et al., 2024a], EAT [Wei et al., 2024], and DARL [Zhang et al., 2025]. On PC-ERM, HPM raises raw class-specific Mahalanobis AUROC from 45.39 to 85.67

on CIFAR-10-LT and from 54.81 to 78.35 on CIFAR-100-LT. It achieves the highest LES (3.08) in the main CIFAR-100-LT comparison, retaining roughly 95% of that comparison's highest AUROC. On ImageNet-100-LT, specialized training retains a representation-level advantage after readout control.

Our contributions are:

- We revisit whether increasingly complex training is necessary for LT-OOD detection by comparing simple and specialized training under controlled detector readouts, and use LES to evaluate OOD performance together with training-time cost.
- We identify two geometric reasons why raw Mahalanobis fails under long-tailed representations: frequency-coupled feature radius and poorly supported tail-class covariance estimation.
- We propose HPM, a post-hoc Mahalanobis repair based on hyperspherical projection and pooled ridge covariance, showing that simple PC-ERM models can achieve strong and efficient OOD detection without auxiliary OOD supervision or retraining.

2 Related Work

Post-hoc OOD detection and Mahalanobis geometry. Classical post-hoc OOD detectors score softmax confidence [Hendrycks and Gimpel, 2017], use input preprocessing or temperature scaling [Liang et al., 2018], or measure feature-space density with Mahalanobis distance [Lee et al., 2018]. Energy scoring [Liu et al., 2020] has become a strong classifier-score baseline, while later detectors exploit feature clipping, nearest neighbors, residual subspaces, or feature compactness [Sun et al., 2021, 2022, Wang et al., 2022a, Sehwag et al., 2021, Ming et al., 2023]. D-KNN separates principal and residual feature subspaces for post-hoc OOD detection [Peng et al., 2026a]. Our work is closest to Mahalanobis-style detection and recent feature-normalized variants [Müller and Hein, 2025]. Mahalanobis++ studies feature normalization for standard, non-long-tailed OOD detection under the conventional shared covariance; with aligned tied-covariance choices, our closed-form score can coincide with its score. We instead study the long-tailed setting, where raw feature radius is frequency-coupled and tail classes cannot support reliable class-specific covariance.

Long-tailed recognition and LT-OOD detection. Long-tailed recognition has been addressed through re-weighting, re-sampling, margin adjustment, logit adjustment, contrastive learning, and decoupled training [Cao et al., 2019, Kang et al., 2020, Menon et al., 2021, Ren et al., 2020, Zhu et al., 2022, Li et al., 2022]. LT-OOD detection adds the open-set requirement that rare ID samples remain accepted while unknown samples are rejected. Recent methods improve this setting with contrastive tail-OOD separation, abstention or calibrated outlier classes, adaptive outlier distributions, gradient-conflict mitigation, and semantic tail prioritization [Wang et al., 2022b, Wei et al., 2024, Miao et al., 2024a,b, Zhang et al., 2025, He et al., 2025]. These approaches mainly change training. We ask how far a stable post-hoc geometry can take simple long-tail training before adding more specialized objectives.

Auxiliary OOD supervision and representation geometry. Outlier exposure can greatly improve OOD detection when suitable auxiliary data are available [Hendrycks et al., 2019], but it also couples OOD rejection with the learned ID geometry. This coupling is especially delicate in long-tailed learning, where head and tail classes receive very different statistical support. In general OOD detection, PID adapts the number of class prototypes during training to capture variation within classes [Peng et al., 2026b]. Our analysis connects LT-OOD behavior to classifier-null scatter, feature radius, covariance spectra, and effective rank, complementing studies of feature geometry and neural collapse [Papyan et al., 2020, Kornblith et al., 2019].

3 Preliminaries

LT-OOD setting. Let $\mathcal{D}_{\text{ID}} = \{(x_i, y_i)\}$ be a long-tailed training set with K ID classes and class counts $\{n_c\}_{c=1}^{K}$. At test time, a detector receives a query x and decides whether it comes from the ID distribution or from an unseen OOD distribution. We use the convention that larger OOD scores indicate samples that are more likely OOD.

Energy score. Given logits $f(x) \in \mathbb{R}^K$ and temperature $T > 0$, the energy score is

$$\mathcal{E}(x) = -T \log \sum_{k=1}^{K} \exp(f_k(x)/T). \tag{1}$$

Classifier row space and null space. Let $h(x) \in \mathbb{R}^d$ be the penultimate feature and let the linear classifier be $f(h) = Wh + b$, where $W = [w_1^\top; \ldots; w_K^\top] \in \mathbb{R}^{K \times d}$. The classifier row space is $\mathcal{R}(W^\top)$, and its orthogonal complement is the classifier-null space $\mathcal{N}(W) = \{v : Wv = 0\}$. We use the orthogonal projectors

$$P_{\text{row}} = W^\top (WW^\top)^\dagger W, \qquad P_{\text{null}} = I - P_{\text{row}}. \tag{2}$$

For any feature $h = P_{\text{row}}h + P_{\text{null}}h$, we have $WP_{\text{null}}h = 0$. Therefore,

$$f(h) = Wh + b = WP_{\text{row}}h + b = f(P_{\text{row}}h), \qquad \mathcal{E}(h) = \mathcal{E}(P_{\text{row}}h). \tag{3}$$

Thus classifier scores, including Energy, are invariant to any perturbation that lies in the classifier-null component $P_{\text{null}}h$.

Classifier-null absolute scatter. To quantify the amount of feature variation that Energy cannot observe, we measure the within-class scatter that lies in the classifier-null subspace. For class c, let

$$\widehat{\mu}_c = \frac{1}{n_c} \sum_{i:y_i=c} h_i, \qquad \widehat{\Sigma}_c = \frac{1}{n_c - 1} \sum_{i:y_i=c} (h_i - \widehat{\mu}_c)(h_i - \widehat{\mu}_c)^\top. \tag{4}$$

We define the absolute classifier-null scatter as

$$A_c^{\text{null}} = \text{Tr}\Big(P_{\text{null}} \widehat{\Sigma}_c P_{\text{null}}\Big) = \frac{1}{n_c - 1} \sum_{i:y_i=c} \|P_{\text{null}}(h_i - \widehat{\mu}_c)\|_2^2 . \tag{5}$$

This quantity is an absolute trace rather than a fraction. It measures how much class-conditional variation is hidden from classifier scores, not just what proportion of the total scatter lies in the null space. For a group of classes $\mathcal{G}$, such as head or tail classes, we report

$$A_{\mathcal{G}}^{\text{null}} = \frac{1}{|\mathcal{G}|} \sum_{c \in \mathcal{G}} A_c^{\text{null}}. \tag{6}$$

4 Diagnosing and Repairing Long-Tailed Mahalanobis Geometry

4.1 Energy has a classifier-null blind spot

Let $h(x) \in \mathbb{R}^d$, logits $f(x) = Wh(x) + b$, and

$$\mathcal{E}(x) = -T \log \sum_{c=1}^{K} \exp(f_c(x)/T). \tag{7}$$

For any $\delta \in \text{null}(W)$, $W(h + \delta) + b = Wh + b$. Thus Energy observes only classifier-visible coordinates. It can be strong when OOD evidence is classifier-aligned, but it is blind to classifier-null feature variation.

4.2 Raw Mahalanobis mixes feature radius with geometry

For raw features $h = ru$, $r = \|h\|_2$, and $\|u\|_2 = 1$,

$$\text{Cov}(h) = \mathbb{E}[r^2 uu^\top] - \mathbb{E}[ru]\mathbb{E}[ru]^\top. \tag{8}$$

When r is coupled with class frequency, raw Mahalanobis mixes semantic displacement with a long-tail nuisance coordinate. We score in hyperspherical coordinates $z = h/\|h\|_2$ to remove this radius term before computing distances.

4.3 Tail classes cannot support class-specific covariance

A class-specific covariance in d dimensions has

$$\operatorname{rank}(\widehat{\Sigma}_c) \le n_c - 1. \tag{9}$$

For tail classes, $n_c \ll d$ is common, so full class-specific precision estimates have little support. HPM keeps class means specific but pools second-order geometry.

4.4 Why hyperspherical projection and pooled covariance repair the metric

Lemma 1 (Pooled covariance expands supported directions)**.** Assume $n_c \ge 2$ for each class. Let

$$\widehat{\mu}_c^z = \frac{1}{n_c} \sum_{i:y_i=c} z_i, \qquad \widehat{\Sigma}_c^z = \frac{1}{n_c - 1} \sum_{i:y_i=c} (z_i - \widehat{\mu}_c^z)(z_i - \widehat{\mu}_c^z)^\top, \tag{10}$$

let $S_c = \operatorname{span}\{z_i - \widehat{\mu}_c^z : y_i = c\}$, and let

$$\widehat{\Sigma}_{\text{pool}} = \sum_{c=1}^{K} \alpha_c \widehat{\Sigma}_c^z, \qquad \alpha_c > 0, \qquad \sum_c \alpha_c = 1. \tag{11}$$

Then

$$\operatorname{range}(\widehat{\Sigma}_{\text{pool}}) = \operatorname{span}\Big(\bigcup_{c=1}^{K} S_c \Big). \tag{12}$$

With ridge regularization, $\widehat{\Sigma}_{\text{pool}} + \lambda I \succ 0$ for any $\lambda > 0$.

Theorem 1 (HPM as stabilized Mahalanobis geometry)**.** Assume $N > K$ and $n_c \ge 2$ for each class. Let $z = h/\|h\|_2$ and define

$$\widehat{\Sigma}_{\text{HPM}} = \frac{1}{N - K} \sum_{c=1}^{K} \sum_{i:y_i=c} (z_i - \widehat{\mu}_c^z)(z_i - \widehat{\mu}_c^z)^\top + \lambda I, \qquad \lambda > 0. \tag{13}$$

Then HPM uses normalized class means as semantic anchors, removes raw feature radius before covariance estimation, and replaces tail-specific covariance with pooled residual geometry. Lemma 1 characterizes the covariance support before ridge regularization; after adding λI, the precision is full-rank, with residual-supported directions shaped by the pooled covariance and orthogonal directions controlled only by the isotropic ridge.

Proofs are given in Appendix A. We use $\lambda = \frac{N}{N-K}\lambda_0$ with reference ridge $\lambda_0 = 10^{-5}$, scaling the covariance and ridge together under the degrees-of-freedom convention.

4.5 The repaired detector

Given the normalized class means and pooled ridge covariance above, the OOD score is

$$s_{\text{HPM}}(x) = \min_c \left(\frac{h(x)}{\|h(x)\|_2} - \widehat{\mu}_c^z \right)^\top \widehat{\Sigma}_{\text{HPM}}^{-1} \left(\frac{h(x)}{\|h(x)\|_2} - \widehat{\mu}_c^z \right). \tag{14}$$

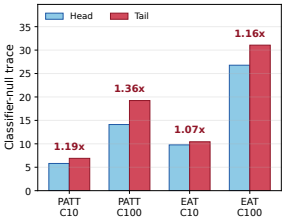
5 Experiments

5.1 Setup

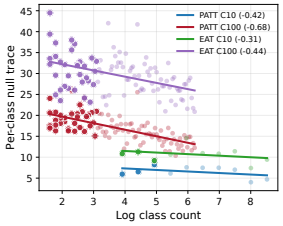
Datasets and models. We evaluate on CIFAR-10-LT, CIFAR-100-LT [Krizhevsky, 2009], and ImageNet-100-LT [Liu et al., 2019, Russakovsky et al., 2015]. The CIFAR-LT OOD suite follows common OOD benchmarks [Yang et al., 2022], covering digit [Netzer et al., 2011], texture [Cimpoi et al., 2014], scene [Yu et al., 2015], and object-level shifts; ImageNet-100-LT uses held-out ImageNet classes as near-OOD samples. CIFAR baselines include PATT [He et al., 2025], PASCL [Wang et al., 2022b], DARL [Zhang et al., 2025], AdaptOD [Miao et al., 2024b], COCL [Miao et al., 2024a], EAT [Wei et al., 2024], and PC-ERM. For methods originally trained with outlier exposure (OE), we report both the original checkpoint and the matched no-OE variant when available. PC-ERM uses the same backbone family and 200-epoch CIFAR budget, but only standard cross-entropy plus long-tail log-prior adjustment [Menon et al., 2021]; it has no auxiliary OOD data, outlier head, contrastive OOD objective, or detector-specific training loss.

Algorithm 1 HPM post-hoc OOD scoring. The algorithm normalizes frozen ID and query features, estimates class anchors with a pooled ridge covariance, and returns a Mahalanobis score where larger values indicate more OOD-like samples.

Require: Frozen encoder $h(\cdot)$; ID training set $\{(x_i, y_i)\}_{i=1}^{N}$ with labels $y_i \in \{1, \ldots, K\}$; query sample x; ridge penalty $\lambda > 0$.
Ensure: Post-hoc OOD score $s_{\mathrm{HPM}}(x)$ (larger implies more OOD-like).
1: Collect training features $\{h_i\}_{i=1}^{N}$ with $h_i \leftarrow h(x_i)$ and the query feature $h \leftarrow h(x)$.
2: Project onto the hypersphere $z_i \leftarrow h_i/\|h_i\|_2$ and $z \leftarrow h/\|h\|_2$.
3: Estimate class anchors $\{\widehat{\mu}_c^z\}_{c=1}^{K}$ and pooled ridge covariance $\widehat{\Sigma}_{\mathrm{HPM}}$ exactly as in Theorem 1.
4: **return** $s_{\mathrm{HPM}}(x) = \min_{c\in[K]} (z - \widehat{\mu}_c^z)^\top \widehat{\Sigma}_{\mathrm{HPM}}^{-1} (z - \widehat{\mu}_c^z)$.



(a) Head–tail comparison



(b) Per-class trend

Figure 2: Classifier-null scatter in CIFAR-LT.

Table 1: **Mahalanobis variants.** Feature normalization and covariance estimation choices for the four post-hoc distance scores.

Name	Feature	Covariance
CS-MD	raw	class-specific
RP-MD	raw	pooled
HC-MD	hyperspherical	class-specific
HPM	hyperspherical	pooled

Scores and unified post-hoc protocol. We compare Energy, maximum softmax probability (MSP), classwise Mahalanobis (CS-MD), raw pooled/tied Mahalanobis (RP-MD), hyperspherical class-specific Mahalanobis (HC-MD), and HPM under a common post-hoc OOD averaging protocol. For CIFAR-LT, we average over the OOD sets listed above. OOD detection is reported with area under the receiver operating characteristic (AUROC) and false positive rate at 95% true positive rate (FPR95); higher AUROC and lower FPR95 are better. ACC denotes closed-set ID classification accuracy. We also report LES,

$$\mathrm{LES} = \lg\left(\frac{\mathrm{BestAUROC}}{\mathrm{Cost}}\right). \tag{15}$$

where BestAUROC is the strongest AUROC in the row and Cost is the total wall-clock time for the complete 200-epoch training run divided by 200. All methods are measured on the same hardware under the same timing rule.

Mahalanobis variants. Let μ_c be the feature mean of class c. A Mahalanobis detector scores a query by

$$s(x) = \min_c (z(x) - \mu_c)^\top \Sigma_c^{-1} (z(x) - \mu_c), \tag{16}$$

where $z(x)$ may be the raw feature $h(x)$ or its normalized version $h(x)/\|h(x)\|_2$, and Σ_c may be class-specific or pooled across classes. Table 1 summarizes the four post-hoc variants. Our method is HPM, which combines hyperspherical projection with pooled covariance.

5.2 Classifier-null blind spot

The following diagnostics explain why the same frozen representation can lead to different conclusions under classifier-score and feature-space detectors. Energy is computed only from classifier logits, so it cannot respond to variation orthogonal to the classifier row space. Figure 2 measures this blind spot in CIFAR-LT checkpoints: tail classes often carry larger classifier-null scatter than head classes, and the null-space component varies systematically with class count.

This result does not say that classifier scores are weak. Instead, it identifies a channel they cannot read. When OOD evidence is aligned with the classifier, Energy or MSP can be strong; when evidence lies in classifier-null variation, a feature-space detector has access to information that classifier outputs discard.

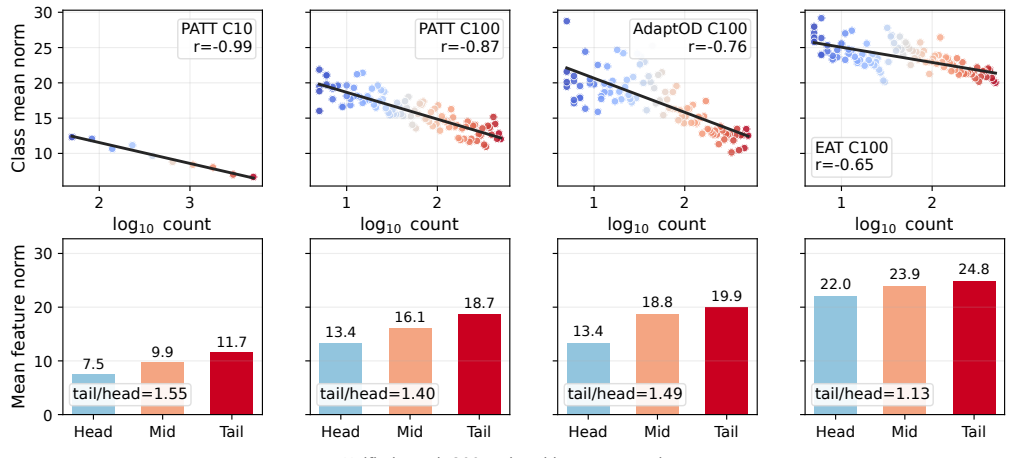


Figure 3: **Feature radius tracks class frequency.** Per-class backbone-feature norm versus training frequency across representative CIFAR-LT checkpoints used in the main tables. Larger norms for lower-frequency classes indicate that raw covariance is fit in coordinates where radius is coupled to imbalance, motivating the hyperspherical projection used before scoring.

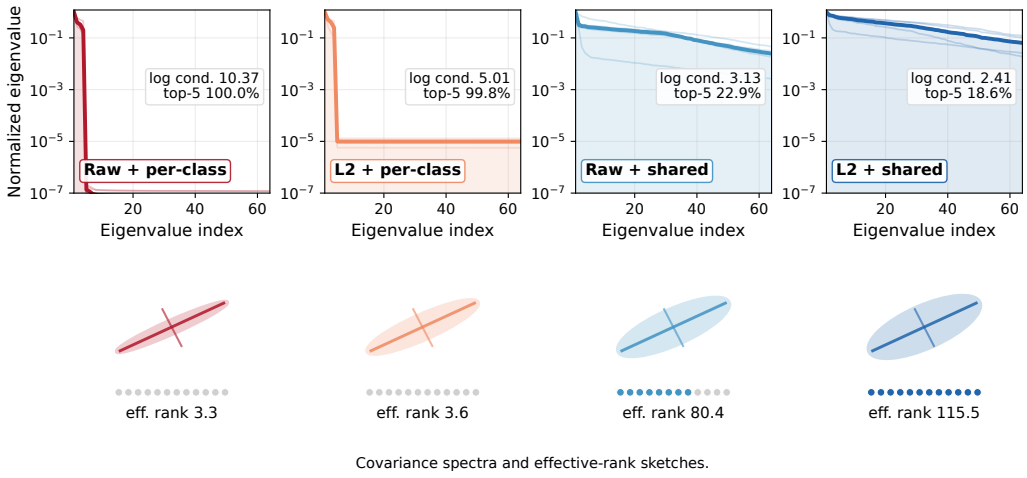


Figure 4: **Pooled hyperspherical covariance is better supported.** Four Mahalanobis variants on CIFAR-100-LT feature banks. Top: eigenvalues of the quadratic form, where sharp drops indicate dependence on a few dominant axes. Bottom: elliptical level sets with effective-rank and conditioning summaries; higher effective rank and lower condition number indicate better-supported geometry.

5.3 Feature-radius coupling

Figure 3 shows the radius-frequency coupling behind hyperspherical projection. In the plotted checkpoints, raw feature norms increase from head to tail groups; the tail/head ratio ranges from about 1.13 to 1.55. Thus raw Mahalanobis can treat class-frequency-induced radius variation as part of the distance geometry.

This explains why feature-space detection still needs repair. A larger radius for rare classes is not necessarily semantic OOD evidence; it can be a training artifact of long-tailed supervision. Hyperspherical projection removes this radial degree of freedom before covariance estimation, forcing the score to compare angular residual structure rather than raw norm scale.

Table 2: **CIFAR-10-LT post-hoc OOD results.** Entries are averaged AUROC / FPR95; detector columns mark row-wise best/second-best values, ACC/LES are column-wise, and only top-three HPM cells are shaded.

Model	Energy	MSP	CS-MD	RP-MD	HC-MD	HPM	ACC	LES↑
Models trained with auxiliary outlier exposure.								
PATT	81.78/51.29	**86.25/46.39**	46.61/87.99	56.80/76.80	82.76/47.57	85.20/44.28	79.86	1.96
AdaptOD	66.72/72.49	67.36/74.30	55.07/84.78	58.98/86.13	72.97/68.10	**77.75/62.88**	57.57	2.84
COCL	**89.75/42.17**	88.21/46.08	62.31/65.99	89.04/42.44	87.75/43.32	89.20/41.93	80.08	2.84
EAT	78.29/59.07	**79.57/60.21**	61.31/83.54	60.40/83.07	74.14/66.20	70.22/71.57	79.50	1.83
Models trained without auxiliary outlier exposure.								
PATT	**90.17/36.17**	87.47/40.06	58.28/76.70	55.30/78.17	85.44/43.34	86.78/41.05	83.01	1.97
PASCL	87.08/42.48	83.53/46.15	66.01/76.59	66.09/76.80	82.38/49.27	**88.89/40.29**	85.78	2.40
DARL	77.20/66.26	83.78/54.22	34.28/90.57	55.20/86.60	76.39/53.49	**88.72/38.15**	**87.71**	2.48
AdaptOD	73.87/64.66	69.48/67.98	66.98/77.81	57.79/81.70	78.52/61.51	**79.36/59.54**	66.58	2.87
COCL	**76.38/59.63**	74.15/61.11	72.04/65.71	59.36/76.31	72.86/60.71	70.58/67.03	76.81	2.79
EAT	65.39/80.19	64.34/79.98	66.50/82.82	55.82/82.95	**68.16/74.35**	65.58/78.49	73.40	1.77
PC-ERM	85.06/50.50	81.32/50.56	45.39/90.61	46.49/92.15	77.65/53.60	**85.67/46.38**	84.09	**3.05**

Table 3: **CIFAR-100-LT post-hoc OOD results.** Same protocol and notation as Table 2.

Model	Energy	MSP	CS-MD	RP-MD	HC-MD	HPM	ACC	LES↑
Models trained with auxiliary outlier exposure.								
PATT	73.86/57.89	78.73/53.96	65.30/70.76	44.13/90.77	75.72/58.08	**81.41/46.81**	48.22	2.03
AdaptOD	**67.22/73.76**	59.85/84.38	43.84/92.25	44.55/91.80	57.24/83.94	51.42/89.33	32.98	2.80
COCL	**73.46/68.85**	71.03/72.49	44.98/93.28	65.93/75.57	68.91/71.93	71.40/69.95	45.10	2.77
EAT	**68.04/75.26**	62.97/82.20	57.19/89.57	58.03/85.57	62.92/81.10	63.39/83.81	43.70	1.80
Models trained without auxiliary outlier exposure.								
PATT	80.29/52.72	79.75/55.40	70.79/71.07	41.33/92.46	76.13/60.38	**82.61/45.56**	51.33	2.03
PASCL	76.20/64.42	70.74/72.87	55.71/88.75	54.16/89.30	73.64/64.55	**79.30/58.46**	50.88	2.43
DARL	62.17/78.50	68.54/74.88	53.95/90.93	47.33/93.39	69.64/71.15	**79.31/59.35**	**56.63**	2.25
AdaptOD	70.96/71.68	65.20/78.56	61.27/82.52	40.59/95.10	69.02/74.84	**72.82/69.86**	35.86	2.86
COCL	70.04/72.35	66.22/77.32	42.21/93.50	52.29/84.81	69.36/74.41	**76.86/66.57**	45.99	2.82
EAT	63.60/81.13	61.76/83.14	60.77/85.63	52.55/89.40	65.50/76.72	**70.09/75.21**	41.39	1.82
PC-ERM	73.44/67.64	67.96/73.05	54.81/90.77	50.40/91.26	71.95/66.46	**78.35/56.92**	51.91	**3.08**

5.4 Covariance distortion

Figure 4 shows the second failure mode: raw per-class covariance is dominated by a few unstable directions. Across PATT, PC-ERM, DARL, and PASCL on CIFAR-100-LT, the median effective rank rises from about 3.3 for raw per-class covariance to about 115.5 for pooled hyperspherical covariance, while the median log-condition number drops from about 10.4 to about 2.4.

The comparison isolates the role of the two design choices in Table 1. Hyperspherical projection removes the radial nuisance identified in Figure 3; pooling then replaces tail-specific covariance estimates with residual directions shared across classes. This is consistent with Lemma 1: pooling does not create new features, but it aggregates complementary residual support; empirically, this produces a better-conditioned metric in Figure 4.

Outlier exposure as a side analysis. We treat OE as a side analysis rather than as the main target. Controlled no-OE/light-OE comparisons show that OE changes ID compactness, classifier-null variation, and radius-covariance coupling. Thus an OOD objective can improve a task-specific score without necessarily producing cleaner post-hoc geometry; this motivates reporting detector behavior on the same frozen checkpoints in the main tables.

5.5 CIFAR-LT results

Tables 2 and 3 compare six post-hoc scores on the same frozen CIFAR-LT representations. This tests when HPM repairs raw Mahalanobis scoring and when the repaired feature-space detector remains competitive with classifier scores.

Result analysis. Among the six main-table detectors, HPM has the highest AUROC for PC-ERM on both datasets and for all no-OE CIFAR-100-LT rows. For PC-ERM, it raises raw CS-MD AUROC from 45.39 to 85.67 on CIFAR-10-LT and from 54.81 to 78.35 on CIFAR-100-LT. Pooling after normalization adds 8.02 and 6.40 points over HC-MD. Energy and MSP remain stronger in some CIFAR-10-LT or OE-trained cases. Figure 5 reports LES for the main-table comparison.

5.6 ImageNet-100-LT near-OOD analysis

Table 4 compares raw CS-MD with HPM on ImageNet-100-LT. Raw classwise Mahalanobis remains consistently weak, while HPM substantially repairs feature-space scoring. Specialized training nevertheless retains a representation-level advantage after readout control, so detector stabilization does not close the gap in every near-OOD setting.

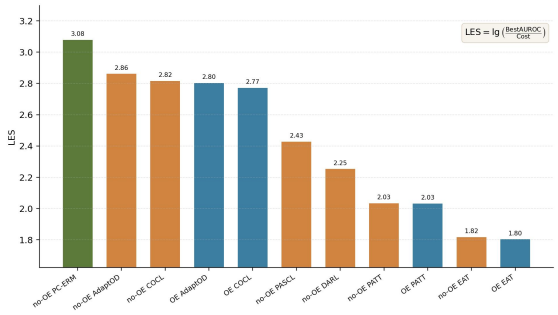


Figure 5: **CIFAR-100-LT efficiency.** LES compares best AUROC with average per-epoch training time; higher is better.

Table 4: **ImageNet-100-LT near-OOD.** Main AUROC/FPR95 comparison of CS-MD and HPM.

Model	CS-MD	HPM	ACC	LES↑
Models trained with auxiliary outlier exposure.				
PATT	46.73/93.03	**81.04/54.33**	59.70	0.97
COCL	41.74/90.34	**81.45/59.35**	42.45	1.96
EAT	57.36/91.26	**80.18/63.43**	39.35	1.67
Models trained without auxiliary outlier exposure.				
PATT	40.99/95.65	**80.04/54.97**	55.60	0.95
COCL	35.26/95.96	**77.70/60.22**	41.20	1.97
EAT	50.47/95.29	**74.92/66.76**	38.70	1.64
PASCL	43.17/95.96	**88.26/37.88**	**66.30**	1.04
PC-ERM	52.85/89.42	**63.47/84.15**	46.40	**2.42**

6 Discussion and Limitations

Energy is effective when OOD separation is visible to the classifier, while HPM reads evidence that becomes usable after stabilizing feature geometry. LT-OOD performance should therefore separate the representation, the detector, and the cost paid to train the model.

PC-ERM suggests that gains from specialized LT-OOD training should be re-evaluated against simple training plus repaired post-hoc geometry. Auxiliary OOD data, extra heads, contrastive losses, energy regularization, and gradient-conflict mitigation can help, but in several CIFAR comparisons the gap narrows once detector geometry is stabilized. With only cross-entropy, long-tail prior calibration, and HPM, PC-ERM remains competitive on CIFAR-LT and obtains the best LES in both main-table CIFAR comparisons. On CIFAR-100-LT, it retains about 95% of that comparison's highest AUROC. On ImageNet-100-LT, specialized representations retain an advantage after detector repair. These results support evaluating representation quality across multiple readouts when assessing the value of specialized training.

7 Conclusion

Frozen feature geometry strongly affects LT-OOD detection. HPM addresses feature-radius nuisance and limited tail covariance support without auxiliary OOD data or retraining. With PC-ERM, it achieves competitive OOD performance and high LES in the main CIFAR-LT comparisons. Specialized representations retain an advantage on ImageNet-100-LT. These results support assessing representation quality across multiple detectors. Code is available at `https://github.com/NNU-XuanmingChen/LT-OOD-Geometry`.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (22673056 and 92370127). The authors declare no competing interests.

References

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, volume 32, pages 1567–1578, 2019.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3606–3613, 2014.

Yina He, Lei Peng, Yongcun Zhang, Juanjuan Weng, Shaozi Li, and Zhiming Luo. Long-tailed out-of-distribution detection: Prioritizing attention to tail. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(3):3446–3454, 2025. doi: 10.1609/aaai.v39i3.32357.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2656–2666, 2019. doi: 10.1109/CVPR.2019.00277.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, volume 31, pages 7167–7177, 2018.

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S. Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6918, 2022. doi: 10.1109/CVPR52688.2022.00679.

Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475, 2020.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2532–2541, 2019. doi: 10.1109/CVPR.2019.00264.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations*, 2021.

Wenjun Miao, Guansong Pang, Xiao Bai, Tianqi Li, and Jin Zheng. Out-of-distribution detection in long-tailed recognition with calibrated outlier class learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):4216–4224, 2024a. doi: 10.1609/aaai.v38i5.28217.

Wenjun Miao, Guansong Pang, Jin Zheng, and Xiao Bai. Long-tailed out-of-distribution detection via normalized outlier distribution adaptation. In *Advances in Neural Information Processing Systems*, volume 37, pages 132106–132132, 2024b. doi: 10.52202/079017-4199.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *International Conference on Learning Representations*, 2023.

Maximilian Müller and Matthias Hein. Mahalanobis++: Improving OOD detection via feature normalization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 45151–45184. PMLR, 2025.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

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*, 117(40): 24652–24663, 2020. doi: 10.1073/pnas.2015509117.

Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1695, 2023. doi: 10.1109/ICCV51070.2023.00162.

Ningkang Peng, Xiaoqian Peng, Yuhao Zhang, Qianfeng Yu, Feng Xing, Peirong Ma, Xichen Yang, Yi Chen, Tingyu Lu, and Yanhui Gu. Breaking semantic hegemony: Decoupling principal and residual subspaces for generalized OOD detection. *arXiv preprint arXiv:2602.05360*, 2026a.

Ningkang Peng, Qianfeng Yu, Xiaoqian Peng, Linjing Qian, Yafei Liu, Canran Xiao, Xinyu Lu, Tingyu Lu, Zhichao Zheng, and Yanhui Gu. How to achieve prototypical birth and death for OOD detection? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 6466–6475, June 2026b.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, volume 33, pages 4175–4186, 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. SSD: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2021.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, volume 34, pages 144–157, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. ViM: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4911–4920, 2022a. doi: 10.1109/CVPR52688.2022.00487.

Haotao Wang, Aston Zhang, Yi Zhu, Shuai Zheng, Mu Li, Alex J. Smola, and Zhangyang Wang. Partial and asymmetric contrastive learning for out-of-distribution detection in long-tailed recognition. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23446–23458. PMLR, 2022b.

Tong Wei, Bo-Lin Wang, and Min-Ling Zhang. EAT: Towards long-tailed out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15787–15795, 2024. doi: 10.1609/aaai.v38i14.29508.

Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8301–8309, 2021.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 35, pages 32598–32611, 2022. doi: 10.52202/068431-2362.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365, 2015.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2(3):1–32, 2024.

Xuan Zhang, Sinchee Chin, Jing-Hao Xue, Xiaochen Yang, and Wenming Yang. DARL: Mitigating gradient conflicts in long-tailed out-of-distribution learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 6868–6877, 2025. doi: 10.1145/3746027.3755127.

Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6898–6907, 2022. doi: 10.1109/CVPR52688.2022.00678.

A Proofs

A.1 Residual support under pooling

Proof of Lemma 1. For any positive semidefinite matrices $\{A_c\}_{c=1}^K$ and positive weights $\alpha_c > 0$,

$$v^\top \Big(\sum_{c=1}^K \alpha_c A_c \Big) v = \sum_{c=1}^K \alpha_c v^\top A_c v. \tag{17}$$

Each term is nonnegative, so the sum is zero if and only if $v^\top A_c v = 0$ for all c. Since $A_c \succeq 0$, this is equivalent to $A_c v = 0$ for all c. Hence

$$\operatorname{null}\Big(\sum_{c=1}^K \alpha_c A_c \Big) = \bigcap_{c=1}^K \operatorname{null}(A_c). \tag{18}$$

Taking orthogonal complements gives

$$\operatorname{range}\Big(\sum_{c=1}^K \alpha_c A_c \Big) = \operatorname{span}\Big(\bigcup_{c=1}^K \operatorname{range}(A_c) \Big). \tag{19}$$

Taking $A_c = \widehat{\Sigma}_c^z$ gives $\operatorname{range}(A_c) = S_c$. For $\lambda > 0$, the ridge-regularized covariance is positive definite. □

A.2 Hyperspherical pooled geometry

Proof of Theorem 1. For $h \neq 0$, HPM applies the map $h \mapsto z = h/\|h\|_2$. Since $ah/\|ah\|_2 = z$ for every $a > 0$, positive rescaling of any training or query feature leaves the fitted detector and its scores unchanged. The weights $\alpha_c = (n_c - 1)/(N - K)$ are positive and sum to one, giving

$$\widehat{\Sigma}_{\mathrm{HPM}} = \widehat{\Sigma}_{\mathrm{pool}} + \lambda I \tag{20}$$

with $\widehat{\Sigma}_{\mathrm{pool}} = \sum_c \alpha_c \widehat{\Sigma}_c^z$. Lemma 1 identifies the pooled covariance support as the span of the within-class residual directions across classes. Adding λI makes the covariance positive definite. The class anchors $\widehat{\mu}_c^z$ remain class-specific. □

B Detector Comparisons

B.1 Nearest-neighbor and Mahalanobis++ comparisons

KNN [Sun et al., 2022] scores each query by its distance to the 50-th nearest normalized training feature. NNGuide [Park et al., 2023] uses $k = 10$ and a 1% ID reference bank sampled with seed 0. We evaluate log-sum-exp (Energy) and MSP base confidences, negating the final NNGuide confidence for OOD scoring. KNN and NNGuide use the same frozen checkpoints, feature layer, and OOD queries as HPM; hyperparameters are fixed before OOD evaluation.

Table 5: **Nearest-neighbor comparisons.** AUROC/FPR95 percentages for fixed checkpoints, with FPR95 measured at 95% OOD recall. PATT-OE denotes the outlier-exposure checkpoint.

Checkpoint	KNN	NNGuide–Energy	NNGuide–MSP	HPM
PATT-OE C10	84.41/46.59	79.48/53.67	85.59/48.06	85.20/44.28
PATT-OE C100	79.67/53.36	79.01/52.75	79.31/53.09	81.41/46.81
PC-ERM C10	81.88/53.10	84.19/55.98	74.35/68.47	85.67/46.38
PC-ERM C100	75.07/68.74	76.49/65.03	70.40/69.85	78.35/56.92

With matched class anchors, covariance, ridge, numerical precision, inversion routine, and score direction, HPM and Mahalanobis++ [Müller and Hein, 2025] use the same normalized tied-covariance scoring operator $\min_c (z - \mu_c)^\top (C + \lambda I)^\dagger (z - \mu_c)$. The maximum per-sample score difference is zero on PATT-OE C10, PC-ERM C10, and PC-ERM C100.

C Long-Tailed Feature Geometry

Class-specific covariance fits in these diagnostics use Cholesky inversion with pseudoinverse fallback; main CIFAR tables use pseudoinverses.

C.1 Feature radius and class frequency

Write $h = ru$, with $r > 0$ and $\|u\|_2 = 1$. For a linear classifier with prior-adjusted logits $a_c = rw_c^\top u + b_c + \tau \log \pi_c$, the margin condition $a_y - a_j \geq \kappa$ requires

$$r(w_y - w_j)^\top u \geq \kappa - (b_y - b_j) + \tau \log(\pi_j/\pi_y). \tag{21}$$

For a tail class y and head class j, the prior term is positive. With weights and direction fixed and $(w_y - w_j)^\top u$ positive, this inequality gives a lower bound on r. The cross-entropy derivative is

$$\frac{\partial \mathcal{L}}{\partial r} = -\sum_{j \neq y} p_j (w_y - w_j)^\top u, \qquad p_c = \mathrm{softmax}(a)_c. \tag{22}$$

The CIFAR PC-ERM cosine head uses $f_c = s\,\overline{w}_c^\top u$ and has $\partial \mathcal{L}/\partial r = 0$ at fixed u. Feature radii can vary as the shared encoder parameters change.

PATT-OE exhibits strong negative radius–frequency correlations, while the cosine-head PC-ERM checkpoints have weaker coupling (Table 6). For each of the 22 CIFAR main-table checkpoints, we measure coupling as the negative Pearson correlation between log class count and mean sample radius. Across checkpoints, this measure has Spearman correlations of 0.174 ($p = 0.437$) with HPM$-$RP-MD gains and -0.291 ($p = 0.189$) with HC-MD$-$CS-MD gains.

Table 6: **Radius-frequency diagnostics.** Spearman ρ: $\log n_c$ versus mean raw sample norm. Tail/head: ratio of mean class-center norms $\|\widehat{\mu}_c\|_2$. Gains are AUROC points.

Checkpoint / head	Spearman ρ	Tail/head	HC-MD$-$CS-MD
PATT-OE C10 / linear	-0.988	1.651	$+36.15$
PATT-OE C100 / linear	-0.880	1.452	$+10.42$
PC-ERM severe LT / cosine	$+0.195$	0.896	$+23.75$
PC-ERM size-matched / cosine	-0.005	1.022	$+14.38$

Both radius statistics vary non-monotonically with τ when seed 42 and the remaining 200-epoch training settings are fixed.

Table 7: **Prior-calibration strength.** Spearman correlation of log class count with mean sample radius, and the tail/head mean-radius ratio. Each dataset–τ setting uses a PC-ERM checkpoint trained with seed 42.

	CIFAR-10-LT		CIFAR-100-LT	
τ	Spearman ρ	Tail/head	Spearman ρ	Tail/head
0.0	-0.394	0.927	$+0.186$	0.904
0.1	-0.346	0.896	$+0.175$	0.901
0.2	-0.442	0.915	$+0.193$	0.912
0.3	-0.455	0.950	$+0.157$	0.917
0.5	-0.139	0.813	$+0.195$	0.913
1.0	-0.455	0.902	$+0.195$	0.893

C.2 Covariance support and imbalance severity

To isolate covariance support, we resample the features used to fit the covariance while holding the encoder, normalized feature bank, class anchors, and ID/OOD queries fixed. Pooling gain is the HPM$-$HC-MD AUROC difference in percentage points.

On the checkpoint trained with balanced data, pooling gains are negative with uniform support of $m \leq 50$ samples per class. With the total fixed at 10,847 samples, changing the allocation from

balanced to LT increases the gain from 1.68 to 12.34 points. With 500 samples from the original LT checkpoints, changing the allocation from uniform to LT increases the pooling gain by 2.32 points for PC-ERM C100 and 1.21 points for PATT-OE C100.

Table 8: **Covariance-support interventions.** A: a PC-ERM checkpoint trained on balanced CIFAR-100; uniform support contains $100m$ samples. B: original LT checkpoints, with 500 total samples allocated uniformly (5/class) or asymmetrically (2–20/class). Results average 20 paired resamples; $m = 500$ uses the full bank once. Intervals are for paired LT-minus-uniform differences.

A. Balanced-training checkpoint							
Samples/class	5	10	20	50	100	200	500
Pooling gain	-11.75	-6.29	-3.40	-0.64	$+1.46$	$+3.45$	$+1.38$
Matched total of 10,847 samples: balanced $+1.68$; LT $+12.34$.							

B. LT-training checkpoints, fixed total support				
Checkpoint	Uniform	LT	Difference	95% CI
PC-ERM C100	-3.25	-0.93	$+2.32$	$[2.07, 2.57]$
PATT-OE C100	-1.19	$+0.02$	$+1.21$	$[1.12, 1.30]$

We also retrain PC-ERM under five CIFAR-100 frequency profiles with seeds 42–44, fixing the architecture, optimization, augmentation, and use of the epoch-200 checkpoint. Full balance uses 500 examples per class; size-matched balance distributes 10,847 examples as 108 or 109 per class. The LT profiles use minimum-to-maximum count ratios of 0.1, 0.02, and 0.01. Severe LT increases normalization and pooling gains over size-matched balance by 8.88 and 3.46 points, respectively.

Table 9: **Three-seed imbalance-severity sweep.** Mean $\pm$ sample standard deviation. Radius ρ is the Pearson correlation of log class count with mean sample norm; a dash denotes constant support. Tail support is the median over the last third of classes ($d = 512$). Normalization gain is HC-MD$-$CS-MD; pooling gain is HPM$-$HC-MD.

Profile	Radius ρ	Tail n_c/d	Norm. gain	Pooling gain	HPM AUROC
Full balanced	–	0.9766	$+14.06 \pm 0.80$	$+0.55 \pm 1.98$	76.97 ± 2.76
Size-matched balanced	-0.050 ± 0.009	0.2129	$+15.22 \pm 1.57$	$+2.43 \pm 0.66$	77.35 ± 1.32
Mild LT	$+0.216 \pm 0.019$	0.1426	$+24.82 \pm 1.44$	$+7.58 \pm 0.89$	78.70 ± 0.95
Medium LT	$+0.250 \pm 0.018$	0.0361	$+26.01 \pm 3.96$	$+6.22 \pm 0.39$	77.78 ± 2.44
Severe LT	$+0.222 \pm 0.026$	0.0205	$+24.10 \pm 0.94$	$+5.89 \pm 0.64$	76.68 ± 1.52

C.3 Classifier-null complementarity

We project raw features onto orthonormal bases Q_{row} and Q_{null} of the classifier row and null spaces, normalize the projected coordinates, and refit HPM. Full features increase AUROC over row-only features by 8.33 points on PATT-OE C10 and 4.14 on PC-ERM C10; row-only features score 2.19 points higher on PC-ERM C100.

Table 10: **Classifier-null ablation.** AUROC (%) for full, row-only and null-only HPM, and Energy.

Frozen bank	Full	Row-only	Null-only	Energy
PATT-OE C10	85.20	76.87	84.37	81.78
PC-ERM C10	85.67	81.53	83.11	85.06
PC-ERM C100	78.35	80.54	75.11	73.44

D Ridge Selection and Sensitivity

We fixed the reference ridge $\lambda_0 = 10^{-5}$ across datasets and checkpoints without OOD tuning. For $a = N/(N - K)$ and residual outer-product sum S, we use $\lambda = a\lambda_0$ and $\widehat{\Sigma}_{\mathrm{HPM}} = a(S/N + \lambda_0 I)$. Scores scale by a^{-1}, preserving their rankings. Table 11 varies λ_0; all fits use ID training features.

Table 11: **Per-checkpoint ridge sensitivity.** HPM AUROC/FPR95 percentages, averaged over five CIFAR OOD sources or 15 ImageNet OOD subsets. Columns give the reference ridge λ_0, with fixed default 10^{-5}.

Checkpoint	$\lambda_0 = 0$	$\lambda_0 = 10^{-5}$	$\lambda_0 = 10^{-4}$
PATT-OE C10	85.43/43.89	85.20/44.28	85.16/44.38
PATT-OE C100	81.29/47.02	81.41/46.81	82.03/45.98
PC-ERM C10	85.47/47.79	85.67/46.38	85.78/45.57
PC-ERM C100	78.14/57.21	78.35/56.92	79.32/55.64
PC-ERM ImageNet-100-LT	63.37/84.94	63.47/84.15	64.61/80.53

For $C = Q\,\mathrm{diag}(\sigma_1, \ldots, \sigma_d)Q^\top$ and $\delta_c = z - \widehat{\mu}_c^z$, ridge weights the residual directions as

$$d_{c,\lambda}(z) = \sum_{j=1}^{d} \frac{(q_j^\top \delta_c)^2}{\sigma_j + \lambda}. \tag{23}$$

As ridge increases, $\lambda s_{\mathrm{HPM},\lambda}(z)$ approaches $\min_c \|z - \widehat{\mu}_c^z\|_2^2$.

E Additional Evaluation

E.1 ImageNet-100-LT readouts

For the main ImageNet protocol, each checkpoint is evaluated on 15 random subsets of 30,000 OOD images from held-out ImageNet-1K classes, with the ID split fixed. PASCL exceeds PC-ERM by 24.79 AUROC points under HPM, 15.80 under Energy, and 14.06 under MSP, with ID accuracy also higher by 19.90 points.

Table 12: **Additional ImageNet-100-LT classifier scores.** Mean AUROC/FPR95 percentages over 15 OOD subsets. The corresponding CS-MD, HPM, and ID accuracy results appear in Table 4.

Checkpoint	Energy	MSP
PATT-OE	88.33/38.78	85.74/45.90
COCL-OE	83.82/48.91	77.93/67.07
EAT-OE	91.01/41.20	74.87/70.81
PATT (no OE)	82.93/49.03	80.65/55.54
COCL (no OE)	80.81/55.58	74.43/70.95
EAT (no OE)	80.00/59.35	71.19/75.91
PASCL	90.82/29.42	83.63/49.44
PC-ERM	75.02/67.11	69.57/72.96

The geometric readout comparison uses the fixed seed-00 OOD subset and the complete ID training bank. PATT-OE and PC-ERM use 2048-dimensional encoder features. PASCL uses its 1024-dimensional projection for distance scores and the 2048-dimensional encoder output for classification.

Table 13: **ImageNet-100-LT readouts.** AUROC percentages on the fixed seed-00 OOD subset.

Representation	HC-MD	HPM	Cosine	Diagonal	Energy
PC-ERM	59.44	63.47	71.00	71.61	74.99
PATT-OE	74.25	80.98	81.99	84.10	88.21
PASCL	82.62	88.23	87.32	87.55	90.75

E.2 Near and far OOD with long-tailed encoders

We evaluate six frozen LT-trained checkpoints using the OpenOOD v1.5 near/far image lists [Zhang et al., 2024], retaining each checkpoint's deterministic 32×32 transforms and normalization. Near OOD comprises cross-CIFAR and TinyImageNet; far OOD comprises MNIST, SVHN, Texture, and

Places365. Metrics are averaged over sources within each group. Detector parameters are fixed, and detector fitting uses ID training features. HPM exceeds HC-MD in all 12 checkpoint–group pairs, with a mean AUROC gain of 5.24 points.

Table 14: **OpenOOD near/far results.** HPM AUROC/FPR95 percentages and HPM minus HC-MD AUROC points on each frozen representation. For FPR95, ID scores at or above the fifth percentile of OOD scores count as false positives.

		HPM AUROC/FPR95		Pooling gain	
ID	Model	Near	Far	Near	Far
C10	PASCL	84.09/54.09	93.81/25.87	$+9.75$	$+6.85$
C10	PATT-OE	78.27/57.67	89.45/36.97	$+0.94$	$+1.55$
C10	PC-ERM	77.66/63.50	87.71/46.32	$+3.56$	$+4.89$
C100	PASCL	60.97/85.45	84.94/47.44	$+7.21$	$+5.76$
C100	PATT-OE	62.68/79.70	88.33/36.86	$+5.53$	$+6.65$
C100	PC-ERM	62.94/81.56	85.30/47.21	$+5.28$	$+4.94$

E.3 CIFAR-LT results by OOD source

The following tables use the SCOOD semantic ID/OOD partition [Yang et al., 2021].

Table 15: **CIFAR-10-LT results by OOD source.** AUROC / FPR95 percentages. Suite averages and formatting conventions are given in Table 2.

Model	OOD	Energy	MSP	CS-MD	RP-MD	HC-MD	HPM
PATT-OE	CIFAR-100	77.11/70.28	**79.31/66.55**	55.87/88.17	49.61/82.93	75.61/60.57	76.40/61.29
	SVHN	86.53/19.72	**98.59/6.28**	37.75/87.57	70.36/66.07	95.37/23.53	98.50/8.89
	Texture	77.93/56.54	85.81/49.33	56.19/83.07	65.65/69.93	87.63/42.36	**92.45/37.58**
	TinyImageNet	77.95/68.06	**79.83/66.34**	52.67/88.18	49.70/82.64	77.28/59.60	78.27/60.34
	LSUN	**89.37/41.83**	87.71/43.46	30.59/92.96	48.69/82.44	77.89/51.78	80.41/53.29
PASCL	CIFAR-100	**84.03/49.87**	81.77/49.72	56.90/86.74	56.15/87.09	72.45/67.26	82.31/59.49
	SVHN	91.94/30.70	87.93/38.37	71.33/65.38	79.16/61.95	92.76/30.01	**98.01/11.59**
	Texture	84.87/45.95	81.91/47.55	72.54/74.64	82.72/62.96	87.00/45.14	**95.55/23.44**
	TinyImageNet	**85.48/48.81**	82.19/50.37	57.30/86.24	56.60/85.65	74.54/64.17	83.83/55.47
	LSUN	**89.10/37.08**	83.84/44.73	71.98/69.96	55.81/86.36	85.17/39.79	84.78/51.46
DARL	CIFAR-100	72.47/69.94	82.03/58.69	32.36/93.91	50.25/90.23	68.53/64.28	**82.85/52.21**
	SVHN	86.39/51.30	86.60/41.05	44.55/81.79	60.26/82.79	82.06/43.29	**96.66/17.11**
	Texture	78.84/67.58	83.42/56.95	50.72/85.45	70.11/80.24	81.99/50.45	**94.78/26.16**
	TinyImageNet	73.12/69.28	82.16/60.06	29.48/94.30	45.83/90.31	68.86/63.66	**83.20/52.85**
	LSUN	75.17/73.19	84.68/54.36	14.30/97.40	49.56/89.42	80.50/45.78	**86.10/42.43**
AdaptOD-OE	CIFAR-100	67.78/77.01	67.65/76.24	50.15/92.42	53.26/90.39	66.83/79.47	**70.49/76.83**
	SVHN	59.35/66.89	66.10/71.75	73.87/67.94	70.41/74.35	77.22/54.77	**87.42/40.05**
	Texture	72.83/63.84	66.91/72.60	72.34/74.97	68.74/84.12	74.41/70.50	**82.01/62.60**
	TinyImageNet	69.36/77.12	68.08/76.61	49.21/92.20	52.01/90.47	67.66/77.30	**71.16/74.73**
	LSUN	64.27/77.58	68.05/74.31	29.77/96.36	50.47/91.31	**78.73/58.44**	77.68/60.18

Table 15: **CIFAR-10-LT results by OOD source (continued).**

Model	OOD	Energy	MSP	CS-MD	RP-MD	HC-MD	HPM
COCL-OE	CIFAR-100	**82.57/61.95**	81.43/62.37	68.94/67.59	80.25/65.49	79.53/61.39	80.71/62.38
	SVHN	95.87/17.39	93.92/26.58	60.46/57.68	**97.43/11.19**	95.29/20.49	96.73/14.92
	Texture	93.90/28.55	92.91/33.48	69.34/58.24	93.79/30.55	92.65/32.88	**94.35/29.79**
	TinyImageNet	**84.62/61.98**	83.13/62.79	65.07/68.38	82.68/63.88	81.46/60.94	82.88/61.75
	LSUN	**91.77/40.97**	89.65/45.17	47.74/78.07	91.07/41.07	89.82/40.89	91.32/40.79
EAT-OE	CIFAR-100	**75.85/68.84**	74.17/70.21	57.97/88.54	56.69/85.97	66.38/78.11	61.40/81.55
	SVHN	83.87/39.27	**86.73/48.18**	60.30/79.95	48.21/87.09	71.70/62.74	68.01/68.46
	Texture	81.35/55.39	82.27/54.55	74.68/70.57	79.51/70.99	**85.45/54.54**	85.45/57.63
	TinyImageNet	74.19/70.78	**74.31/70.68**	57.74/87.34	54.72/87.03	66.36/77.21	61.16/80.71
	LSUN	76.17/61.09	80.36/57.45	55.83/91.30	62.87/84.27	**80.80/58.40**	75.10/69.50
PC-ERM	CIFAR-100	**82.83/56.74**	79.50/55.85	50.74/92.30	43.35/93.84	71.94/62.79	78.31/59.92
	SVHN	88.34/34.99	85.05/38.44	44.88/85.15	35.27/94.13	83.35/39.01	**95.49/21.47**
	Texture	84.92/51.35	82.88/45.75	61.86/86.19	67.76/85.60	80.99/50.50	**90.82/37.88**
	TinyImageNet	**84.13/55.29**	79.44/58.64	47.16/92.73	42.60/93.39	73.32/62.85	80.02/59.61
	LSUN	**85.08/54.11**	79.71/54.13	22.31/96.70	43.45/93.80	78.66/52.83	83.70/53.02

Table 16: **CIFAR-100-LT results by OOD source.** Same notation as Table 15; suite averages appear in Table 3.

Model	OOD	Energy	MSP	CS-MD	RP-MD	HC-MD	HPM
PATT-OE	CIFAR-10	**64.23/79.96**	64.21/80.24	51.05/94.23	45.99/93.51	55.15/84.90	59.77/81.74
	SVHN	84.04/35.19	90.57/34.49	74.02/64.13	43.21/90.02	89.31/42.88	**95.77/19.33**
	Texture	65.12/78.17	70.28/76.79	54.30/91.97	73.31/79.57	79.62/66.28	**87.57/52.54**
	TinyImageNet	**69.36/77.17**	69.07/76.48	50.29/93.67	48.11/91.90	57.47/83.66	64.15/79.62
	LSUN	86.55/18.97	99.50/1.80	96.85/9.78	10.04/98.85	97.05/12.70	**99.80/0.83**
PASCL	CIFAR-10	63.36/87.39	**64.26/80.77**	46.39/94.94	48.88/92.79	49.55/90.72	55.29/89.73
	SVHN	95.86/22.09	86.01/52.39	59.22/81.13	48.52/89.99	93.16/32.15	**96.13/23.11**
	Texture	62.87/90.68	57.72/91.46	55.95/92.65	78.19/78.87	80.42/67.10	**89.71/52.04**
	TinyImageNet	**69.93/76.84**	66.37/79.58	43.51/95.71	50.31/92.65	56.84/85.67	64.98/82.83
	LSUN	88.97/45.09	79.32/60.15	73.48/79.32	44.91/92.22	88.25/47.13	**90.38/44.61**
DARL	CIFAR-10	67.28/73.47	**67.59/76.78**	43.12/96.01	46.17/94.66	49.50/87.13	58.29/88.57
	SVHN	61.55/72.84	70.82/68.46	59.98/86.34	55.48/90.33	82.90/58.03	**92.19/34.77**
	Texture	44.76/90.57	62.57/83.06	56.45/92.84	76.92/87.24	80.76/69.07	**90.26/48.89**
	TinyImageNet	68.03/76.78	**70.37/75.65**	45.82/94.14	46.30/94.87	54.40/85.96	66.64/78.54
	LSUN	69.22/78.86	71.37/70.44	64.39/85.32	11.76/99.84	80.63/55.58	**89.14/45.97**
AdaptOD-OE	CIFAR-10	56.63/82.95	**59.80/83.88**	51.91/90.87	48.18/91.54	52.61/89.61	55.29/89.73
	SVHN	**72.08/63.76**	58.00/82.00	19.59/98.04	44.88/91.43	66.42/76.77	53.90/85.23
	Texture	**75.17/69.69**	59.46/86.90	50.93/92.55	61.72/86.04	64.75/81.11	61.76/87.09
	TinyImageNet	**61.01/83.85**	60.29/84.78	52.09/90.48	45.83/93.27	50.48/89.06	50.64/91.55
	LSUN	**71.19/68.53**	61.71/84.36	44.67/89.30	22.13/96.74	51.97/83.17	35.52/93.04
COCL-OE	CIFAR-10	**64.54/79.76**	63.49/81.60	53.81/92.69	55.05/86.01	56.12/83.81	59.79/81.65
	SVHN	75.44/59.15	74.11/64.48	34.43/94.67	85.71/57.82	80.16/58.73	**87.34/49.63**
	Texture	78.94/66.19	74.77/70.49	37.12/97.66	82.83/66.53	82.80/65.57	**86.09/61.35**
	TinyImageNet	**69.12/78.62**	67.26/79.20	50.49/92.86	57.37/83.93	59.85/81.99	63.08/80.34
	LSUN	**79.26/60.55**	75.52/66.69	49.04/88.50	48.68/83.54	65.59/69.57	60.73/76.80
EAT-OE	CIFAR-10	**59.01/84.78**	58.29/86.62	45.07/95.22	46.07/94.13	50.03/89.19	48.43/93.09
	SVHN	**73.58/65.21**	64.37/76.83	65.89/79.50	54.67/85.77	56.16/81.78	54.97/86.66
	Texture	69.72/79.64	62.79/85.23	65.50/89.61	83.41/68.04	81.49/73.46	**84.47/72.54**
	TinyImageNet	**60.62/87.27**	59.75/86.62	52.65/93.61	51.19/92.00	54.05/89.26	55.22/91.29
	LSUN	**77.27/59.38**	69.64/75.72	56.83/89.89	54.81/87.91	72.86/71.79	73.87/75.45
PC-ERM	CIFAR-10	**65.13/86.79**	63.92/77.72	46.50/95.71	42.24/96.14	55.81/84.17	59.82/86.74
	SVHN	78.95/48.85	72.08/65.29	43.23/94.83	62.13/88.65	71.62/66.80	**81.40/55.92**
	Texture	70.79/70.97	65.14/78.01	63.76/88.44	82.61/77.23	81.45/64.54	**87.80/50.86**
	TinyImageNet	**71.68/73.44**	66.53/75.42	48.88/95.55	49.63/94.39	57.74/83.52	64.77/79.76
	LSUN	80.64/58.14	72.12/68.79	71.67/79.31	15.40/99.89	93.11/33.26	**97.95/11.30**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and introduction state the paper's scope: post-hoc stabilized Mahalanobis geometry for LT-OOD detection, evaluated on CIFAR-LT and ImageNet-100-LT.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The discussion notes ImageNet-100-LT as a boundary case and separates OOD-head optimization from detector-friendly feature geometry.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes].

Justification: The method section states the geometric setup and assumptions, and the appendix gives the detailed derivations.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: The experiments section specifies the datasets, evaluated models, OOD scores, metrics, and LES definition; the appendix gives ImageNet-100-LT and per-OOD details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes].

Justification: The public repository at `https://github.com/NNU-XuanmingChen/LT-OOD-Geometry` provides the CIFAR-10-LT and CIFAR-100-LT PC-ERM training pipeline, frozen-feature extraction, Energy/MSP/Mahalanobis scoring, compact result snapshots, and run instructions. It does not redistribute datasets, checkpoints, feature banks, ImageNet or third-party baseline repositories, or historical timing/table/plot scripts.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes].

Justification: The experiments section summarizes datasets, auxiliary OOD sources, metrics, and the post-hoc evaluation protocol; the artifact gives runnable commands and hyperparameters.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes].

Justification: Appendix Table 9 reports mean $\pm$ sample standard deviation over seeds 42–44; Table 8 reports paired-resampling confidence intervals. Main-model comparisons use fixed checkpoints.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes].

   Justification: Cost in LES is the total wall-clock time for the complete 200-epoch training run divided by 200.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work uses standard image recognition datasets and post-hoc scoring of learned representations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes].

Justification: Better OOD scoring can improve reliability in long-tailed recognition, but OOD scores should not be used as stand-alone safety guarantees in deployment.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The paper releases no generative models, scraped datasets, or dual-use foundation models.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: The paper cites the datasets and baseline methods used in the experiments. The artifact does not redistribute restricted datasets or full third-party repositories.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes].

    Justification: The public repository at `https://github.com/NNU-XuanmingChen/LT-OOD-Geometry` documents and releases the CIFAR training, feature extraction, scoring, and compact result files described above. Datasets and large artifacts are obtained separately; ImageNet and third-party baseline code are outside this partial release.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing experiments or human-subject studies.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A].

Justification: The paper does not involve crowdsourcing or human-subject research, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: The proposed method, experiments, and scientific claims do not use LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.