MahaVar: OOD Detection via Class-wise Mahalanobis Distance Variance under Neural Collapse

Donghwan Kim
Department of Industrial Engineering
Yonsei University
Seoul, South Korea
cain0709@yonsei.ac.kr

Hyunsoo Yoon[†]
Department of Industrial Engineering
Yonsei University
Seoul, South Korea
hs.yoon@yonsei.ac.kr

Abstract

Out-of-distribution (OOD) detection is a critical component for ensuring the reliability of deep neural networks in safety-critical applications. In this work, we present a key empirical observation: for in-distribution (ID) samples, class-wise Mahalanobis distances exhibit a pronounced sharp minimum structure, where the distance to the nearest class is small while distances to all other classes remain large, resulting in high variance across classes. In contrast, OOD samples tend to exhibit a less pronounced sharp minimum structure, producing comparatively lower variance across classes. We further provide a theoretical analysis grounding this observation in Neural Collapse geometry: under relaxed Neural Collapse assumptions on within-class compactness and inter-class separation, ID samples are shown to structurally exhibit high class-wise distance variance, offering a theoretical basis for its use as an OOD score. Motivated by this observation and its theoretical backing, we propose MahaVar, a simple and effective post-hoc OOD detector that augments the Mahalanobis distance with a class-wise distance variance term. Following the OpenOOD v1.5 benchmark protocol, MahaVar achieves state-of-the-art performance on CIFAR-100 and ImageNet, with consistent improvements in both AUROC and FPR@95 over existing Mahalanobis-based methods across all benchmarks.

1 Introduction

Deploying deep neural networks in real-world environments inevitably raises a critical challenge: models trained under a closed-world assumption often encounter test samples that fall outside their training distribution, leading to overconfident and potentially catastrophic predictions. Out-of-distribution (OOD) detection addresses this challenge by identifying such samples before they cause erroneous decisions, making it an essential component for building reliable and trustworthy AI systems [46]. The importance of OOD detection has been widely recognized across a range of safety-critical domains, including autonomous driving [37], medical imaging [16], and industrial inspection [10], where the consequences of model failure can be severe. OOD detection methods generally build upon pre-trained neural networks and can be broadly categorized into two families. The first family consists of training-based methods [28, 9, 50, 40], which require either access to auxiliary OOD datasets or modification of the training objective to explicitly shape the decision boundary between ID and OOD data. However, such approaches are inherently constrained by the quality and coverage of the available outlier data, and retraining a model for every new deployment scenario is often impractical. The second family consists of post-hoc methods, which leverage the network's representations [39, 36, 20, 29, 32, 23], output logits [14, 24, 43, 25], or activation shaping [38, 45, 6, 44, 17, 12, 11] of a pre-trained classifier without any additional training or access

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

to OOD data. This training-free nature makes post-hoc methods significantly more practical for real-world deployment, where modifying the underlying model is often infeasible.

Among post-hoc methods, Mahalanobis distance-based detectors [20, 29] have demonstrated strong OOD detection performance by computing, for each test sample, the minimum Mahalanobis distance to the class-conditional Gaussian distributions fitted on training data, where each class is characterized by its mean vector and a shared tied covariance matrix. Building upon this framework, we identify a key empirical observation: for ID samples, the sorted class-wise Mahalanobis distances exhibit a pronounced sharp minimum structure, in which the distance to the nearest class is substantially smaller than the distances to all remaining classes. In contrast, OOD samples tend to exhibit a less sharp minimum structure across classes, resulting in lower variance. This structural asymmetry between ID and OOD samples motivates the use of class-wise distance variance as a complementary and discriminative OOD signal.

We explain this phenomenon through the lens of Neural Collapse [31], a geometric phenomenon observed during the terminal phase of deep network training, wherein penultimate features of each class collapse toward their class mean, and the class means arrange themselves into a simplex Equiangular Tight Frame (ETF) with equal inter-class distances and angles. Under relaxed Neural Collapse assumptions, we theoretically analyze the class-wise variance of Mahalanobis distances for ID samples and provide a principled explanation for why the sharp minimum structure emerges. We further discuss conditions under which OOD samples exhibit lower class-wise distance variance, offering a geometric justification for the proposed score.

Motivated by this observation and its theoretical grounding, we propose MahaVar, a simple yet effective post-hoc OOD score that builds upon Mahalanobis++ [29] by augmenting its nearest class Mahalanobis distance with a class-wise distance variance term, addressing the remaining room for improvement in this state-of-the-art framework. The key insight is that by jointly considering both where a sample lies relative to its nearest class and how dispersed its distances are across all classes, MahaVar captures the richer geometric structure of the feature space in a way that neither term alone can achieve. To validate the effectiveness of MahaVar, we conduct extensive experiments following the benchmark protocol proposed in OpenOOD v1.5 [47], across three ID settings including CIFAR-10, CIFAR-100 [18], and ImageNet [4], under both Near-OOD and Far-OOD evaluation settings. MahaVar consistently outperforms all Mahalanobis-based baselines across all three benchmarks, and achieves state-of-the-art performance among all evaluated methods on CIFAR-100 and ImageNet. Notably, on the large-scale ImageNet benchmark averaged over three backbone architectures (ResNet-50 [13], Swin-B [26], and ViT-B [8]), MahaVar achieves the best average AUROC and FPR@95 across all evaluated baselines, demonstrating that the proposed variance term provides consistent gains regardless of the underlying backbone architecture.

In summary, this work makes the following threefold contributions.

- We present a key empirical observation that class-wise Mahalanobis distances exhibit a pronounced sharp minimum structure for in-distribution samples, while OOD samples exhibit a less sharp minimum structure across classes. To our knowledge, this distance-rank structure has not been previously characterized or exploited for OOD detection.

- We provide a **theoretical analysis grounded in Neural Collapse geometry** that characterizes the conditions under which in-distribution samples exhibit strictly higher class-wise distance variance than OOD samples, offering a principled justification for the use of variance as a discriminative OOD signal.

- We propose **MahaVar**, a simple post-hoc OOD detector that augments the standard Mahalanobis score with a class-wise distance variance term. Following the OpenOOD v1.5 benchmark protocol, MahaVar achieves state-of-the-art performance on CIFAR-100 and ImageNet, and consistently outperforms all Mahalanobis-based baselines across all three benchmarks in both AUROC and FPR@95.

2 Related Work

2.1 Post-hoc OOD Detection Methods

Post-hoc methods detect OOD samples using a pre-trained classifier without any additional training or access to OOD data, and can be broadly categorized into representation-based, logit-based, and activation shaping methods.

Representation-based methods leverage the geometry of the penultimate layer features to construct OOD scores. Lee et al. [20] proposed the Mahalanobis distance to class-conditional Gaussian distributions fitted on penultimate layer features, establishing the foundation for distance-based OOD detection. Ren et al. [36] extended this with a relative Mahalanobis distance that normalizes class-conditional distances against a global reference distribution to improve near-OOD detection. Müller and Hein [29] further improved robustness by applying L2 normalization to penultimate layer features prior to distance computation, aligning the feature distributions more closely with the Gaussian assumptions underlying the Mahalanobis distance. To avoid such parametric assumptions, Sun et al. [39] proposed a non-parametric KNN-based score that directly measures distances to training samples in the penultimate layer feature space, and Park et al. [32] built on this by weighting the energy score using cosine similarity to the k-nearest neighbors. Liu and Qin [23] leveraged Neural Collapse geometry to show that centered ID features cluster near their predicted class weight vectors, proposing an angle-based proximity score complemented by feature norm filtering. Liu and Qin [22] proposed fDBD, which measures the distance from the test sample to the classifier's decision boundary. CADRef [21] partitions feature components based on the sign alignment between their weighted contribution and the classifier weight, then normalizes the class-conditional deviation within each partition by a logit-based score.

Logit-based methods derive OOD scores directly from classifier outputs. Hendrycks and Gimpel [14] introduced the maximum softmax probability as a simple yet effective baseline. Building upon this, Liu et al. [24] proposed the energy score, which aggregates all class logits and provides better theoretical grounding than softmax-based scores. Liu et al. [25] further improved sensitivity by applying a generalized entropy over the top-M sorted class probabilities, capturing small entropy variations among the most confident predictions. Wang et al. [43] constructs a virtual OOD class logit from the norm of the feature residual against a principal subspace, and combines it with the original class logits via softmax to produce a unified OOD score.

Activation shaping methods modify penultimate layer activations at inference time to amplify the gap between ID and OOD samples. Sun et al. [38] proposed clipping abnormally large activations in the penultimate layer, observing that OOD samples tend to produce extreme activation patterns. Djurisic et al. [6] extended this idea with ASH, which prunes activations below a threshold and rescales the remaining ones based on per-sample statistics. Xu et al. [45] further generalized activation shaping by applying both lower and upper truncation to suppress abnormally small and large values. Subsequently, Xu et al. [44] showed that scaling alone, without pruning, achieves competitive performance while avoiding the accuracy degradation associated with activation removal. Furthermore, OptFS [48] learns an optimal feature scaling vector directly from the training data, while AdaSCALE [35] dynamically adjusts the percentile threshold per sample based on an OOD likelihood estimated through gradient-based input perturbation.

2.2 Neural Collapse

Neural Collapse is a geometric phenomenon first observed by Papyan et al. [31] during the terminal phase of deep network training, wherein the penultimate layer features and the linear classification head simultaneously converge to a highly structured configuration. Formally, let $\mathbf{h}_c^i$ denote the penultimate layer feature of the i-th training sample with class label c, $\boldsymbol{\mu}_c = \text{Avg}_i\{\mathbf{h}_c^i\}$ the class mean, $\boldsymbol{\mu}_G = \text{Avg}_{i,c}\{\mathbf{h}_c^i\}$ the global mean, and $\mathbf{w}_c$ the weight vector of class c in the linear classification head. Neural Collapse comprises four inter-related properties:

- **NC1 (Within-class variability collapse):** The within-class covariance $\Sigma_W = \text{Avg}_{i,c}\{(\mathbf{h}_c^i - \boldsymbol{\mu}_c)(\mathbf{h}_c^i - \boldsymbol{\mu}_c)^\top\} \to 0$, meaning that features of samples within the same class collapse toward their class mean.

- **NC2 (Convergence to a simplex ETF):** The centered class means $\{\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\}$ converge to a simplex Equiangular Tight Frame (ETF), satisfying

$$\frac{(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)^\top (\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G)}{\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2 \|\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G\|_2} \to \frac{C}{C-1}\delta_{c,c'} - \frac{1}{C-1}, \tag{1}$$

  where $\delta_{c,c'}$ is the Kronecker delta. This implies that all centered class means $\{\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\}$ become equinorm and maximally separated in angle, forming the most spread-out configuration achievable for C vectors in high-dimensional space.

- **NC3 (Convergence to self-duality):** The classifier weight vectors align with the centered class means up to a scaling factor:

$$\frac{\mathbf{w}_c}{\|\mathbf{w}_c\|_2} - \frac{\boldsymbol{\mu}_c - \boldsymbol{\mu}_G}{\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2} \to 0, \tag{2}$$

  establishing a dual relationship between the feature geometry and the classification head.

- **NC4 (Simplification to nearest class center):** The classifier simplifies to a nearest class center decision rule:

$$\underset{c\in\{1,\ldots,C\}}{\operatorname{argmax}} \ \mathbf{w}_c^\top \mathbf{h} + b_c \to \underset{c\in\{1,\ldots,C\}}{\operatorname{argmin}} \ \|\mathbf{h} - \boldsymbol{\mu}_c\|_2, \tag{3}$$

  meaning that classification based on the learned linear head becomes equivalent to assigning the sample to its nearest class mean.

While complete Neural Collapse requires prolonged training beyond zero training error, its convergence trends and corresponding geometric patterns have been consistently observed across diverse architectures and classification tasks before full convergence [23]. This practical prevalence has motivated its application to OOD detection [23, 1], where the structured geometry of ID features serves as a reference for detecting anomalous samples.

3 Sharp Minimum Structure of Class-wise Mahalanobis Distances

3.1 Mahalanobis Distance-based OOD Detection

The Mahalanobis distance-based detector [20] leverages class-conditional statistics estimated from the penultimate layer features of the training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$. For each class c, the class mean $\hat{\boldsymbol{\mu}}_c$ is computed over the penultimate features $\phi(\cdot)$ of all training samples belonging to that class. A single covariance matrix $\hat{\Sigma}$ is shared across all classes and estimated by pooling the deviations of the penultimate features from their respective class means:

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c} \sum_{i:y_i=c} \phi(\mathbf{x}_i), \qquad \hat{\Sigma} = \frac{1}{N} \sum_{c=1}^{C} \sum_{i:y_i=c} \left(\phi(\mathbf{x}_i) - \hat{\boldsymbol{\mu}}_c\right) \left(\phi(\mathbf{x}_i) - \hat{\boldsymbol{\mu}}_c\right)^\top \tag{4}$$

where N_c and N denote the number of samples in class c and in total, respectively. For a test sample $\mathbf{x}$, the squared Mahalanobis distance between its penultimate feature $\phi(\mathbf{x})$ and each class mean is computed as:

$$d_{\text{Maha},c}(\mathbf{x}) = \left(\phi(\mathbf{x}) - \hat{\boldsymbol{\mu}}_c\right)^\top \hat{\Sigma}^{-1} \left(\phi(\mathbf{x}) - \hat{\boldsymbol{\mu}}_c\right) \tag{5}$$

This formulation rests on the assumption that the penultimate features of each class follow a class-conditional Gaussian distribution $\mathcal{N}(\hat{\boldsymbol{\mu}}_c, \hat{\Sigma})$ with a shared tied covariance matrix $\hat{\Sigma}$ across all classes. Intuitively, a test sample is more likely to be in-distribution if its penultimate feature is close to at least one class mean in the Mahalanobis sense. This motivates the following OOD score:

$$S(\mathbf{x}) = -\min_{c\in\{1,\ldots,C\}} d_{\text{Maha},c}(\mathbf{x}) \tag{6}$$

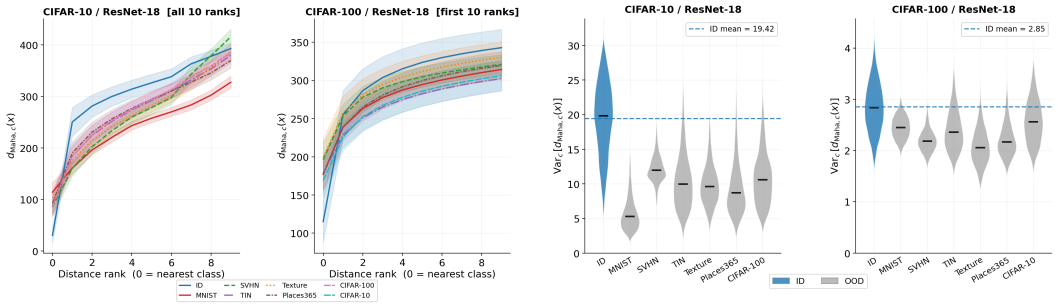


Figure 1: Sorted class-wise $d_{\text{Maha},c}(\mathbf{x})$ (left two panels) and the corresponding distribution of $\text{Var}_c[d_{\text{Maha},c}(\mathbf{x})]$ for ID and OOD samples (right two panels) of Mahalanobis++ [29] on CIFAR-10 and CIFAR-100 as ID datasets, evaluated on OOD datasets following the OpenOOD v1.5 benchmark protocol [47] with ResNet-18 as the backbone. In the sorted distance plots, the x-axis represents the Mahalanobis distance rank, where rank 0 corresponds to the nearest class mean, and shaded regions indicate $\pm 0.5\sigma$ across samples at each rank. In the variance distribution plots, the dashed blue line indicates the mean variance of ID samples, and black bars indicate the median of each distribution.

where a higher score indicates a greater likelihood of the sample being in-distribution. A test sample is classified as OOD when $S(\mathbf{x})$ falls below a threshold T, which is typically calibrated by fixing the true positive rate (TPR) at 95% on the in-distribution validation set.

Müller and Hein [29] proposed Mahalanobis++, which further improves upon this formulation by applying L2 normalization to the penultimate features prior to computing the Mahalanobis distance, encouraging the normalized features to better satisfy the class-conditional Gaussian assumption. In this work, we build upon this L2-normalized Mahalanobis framework and demonstrate that there remains further room for improvement. Specifically, we observe that the class-wise Mahalanobis distances exhibit a sharp minimum structure that carries discriminative information beyond what the minimum distance alone captures, motivating our proposed method.

3.2 Observation: Sharp Minimum in Class-wise Mahalanobis Distances

To motivate our method, we conduct an empirical analysis using Mahalanobis++ under the OpenOOD v1.5 benchmark protocol. Specifically, we consider CIFAR-10 and CIFAR-100 as ID datasets, each with six OOD datasets. For each ID setting, we use a ResNet-18 pretrained on the respective ID dataset and extract the L2-normalized penultimate features. The class-wise means $\hat{\boldsymbol{\mu}}_c$ and tied covariance matrix $\hat{\Sigma}$ are then estimated following Equation 4. For each test sample, we compute the Mahalanobis distance to all class means and sort them in ascending order.

As shown in the left two panels of Figure 1, which plot the sorted class-wise Mahalanobis distances for ID and OOD samples, the results reveal a consistent pattern: ID samples exhibit a substantially lower Mahalanobis distance at rank 0 compared to OOD samples, followed by a sharp increase that quickly plateaus at higher ranks. This indicates that ID samples tend to lie close to a specific class mean while remaining far from all others, whereas OOD samples maintain comparatively large distances to all class means without a clear dominant minimum.

Furthermore, this sharp minimum structure directly implies elevated class-wise distance variance for ID samples: one distance is substantially smaller while the rest plateau at similarly large values, yielding high variance, whereas the less sharp structure of OOD samples yields comparatively lower variance, as verified in the right two panels of Figure 1. There, OOD samples consistently exhibit lower class-wise distance variance than ID samples across all OOD datasets on both CIFAR-10 and CIFAR-100, with the ID mean variance substantially exceeding the median of every OOD distribution. This separation suggests that $\text{Var}_c[d_{\text{Maha},c}(\mathbf{x})]$ naturally captures the class-wise distance structure of ID samples, making it a reliable complementary OOD signal alongside the minimum distance term. Additional results for the original Mahalanobis detector [20] on CIFAR-10/100 and ImageNet, as well as Mahalanobis++ on ImageNet, are provided in Appendix A.

While existing Mahalanobis-based methods either focus solely on the nearest class distance [20, 29] or consider a global distributional perspective [36], the observed class-wise distance structure suggests

unexploited discriminative information, motivating the theoretical analysis and method proposed in the following section.

4 Neural Collapse Perspective on ID/OOD Variance Separation

In this section, we provide a theoretical analysis of the class-wise distance variance structure under the framework of Neural Collapse [31]. As introduced in Section 2.2, Neural Collapse describes a set of geometric properties that emerge in deep neural networks during the terminal phase of training. Notably, Neural Collapse tendencies have been observed even before full convergence, suggesting that its geometric structure is a general property of well-trained classifiers rather than a phenomenon exclusive to the terminal phase.

However, perfect Neural Collapse is rarely achieved in practice, as it requires strict convergence conditions that are difficult to satisfy with finite data and standard training procedures. We therefore adopt a relaxed formulation, retaining only the geometric properties pertaining to the penultimate feature space while excluding assumptions on the classifier weights. Specifically, we discard NC3 (convergence to self-duality) and NC4 (simplification to nearest class-center), as these involve the classifier weight vectors, which are not accessible in a post-hoc detection setting and are not required for our analysis. Instead, we rely solely on weakened versions of NC1 and NC2, which characterize the geometry of the penultimate features and class means, respectively. We formalize these as the following:

Assumption 4.1 (Weak NC1 and NC2)**.** *Let $\phi : \mathcal{X} \to \mathbb{R}^d$ denote the penultimate feature extractor of a pretrained network. We make the following two assumptions:*

(Weak NC1) *For each class $c \in \{1, \dots, C\}$, all $\boldsymbol{x}_i \in \mathcal{X}_{in}$ with $y_i = c$ lie within an ℓ_2-ball of radius ϵ_c centered at the class mean $\boldsymbol{\mu}_c$:*

$$\|\phi(\boldsymbol{x}_i) - \boldsymbol{\mu}_c\|_2 \leq \epsilon_c \quad \forall i : y_i = c \tag{7}$$

This relaxes the strict NC1 condition, which requires within-class variability to vanish (i.e., $\epsilon_c \to 0$), by permitting a bounded deviation of radius $\epsilon_c > 0$.

(NC2) *The centered class means $\{\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\}_{c=1}^{C}$ form a simplex ETF under the ℓ_2-distance, i.e., they are equinorm with $\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2 = R$ for all c, and equiangular with inter-class distance K:*

$$\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|_2 = K \quad \forall c \neq c' \tag{8}$$

Furthermore, we assume $\epsilon = \max_c \epsilon_c < K/2$.

Assumption 4.1 states that the class means form a simplex ETF, while allowing each in-distribution sample to deviate from its class mean by at most ϵ_c in the ℓ_2 sense. The constraint $\epsilon < K/2$ ensures that each sample remains closer to its own class mean than to any other class mean, preserving the nearest-class assignment implied by the ETF structure. Note that a simplex ETF with C classes requires a penultimate layer dimension of at least $C - 1$. For ImageNet ($C = 1000$), this condition is not satisfied by Transformer-based backbones such as ViT-B ($d = 768$), which prevents the class means from forming a complete simplex ETF and consequently weakens the variance separation between ID and OOD samples observed in Appendix A. Under this assumption, we establish bounds on the class-wise ℓ_2 distance variance between ID samples and their class means in Theorem 4.2.

Theorem 4.2 (Bounds on ID Class-wise ℓ_2 Distance Variance)**.** *Let $\phi : \mathcal{X} \to \mathbb{R}^d$ be the penultimate feature extractor of a pretrained network, where $\mathcal{X}_{in} \subset \mathcal{X}$ denotes the in-distribution input space. Suppose Assumption 4.1 holds for ϕ on all $\mathbf{x} \in \mathcal{X}_{in}$. Then for any test ID sample $\mathbf{x} \in \mathcal{X}_{in}$, the class-wise squared ℓ_2 distance variance satisfies:*

$$\frac{(C-1)K^4}{C^2}(1-\gamma)^2 \leq \mathrm{Var}_c\left[\|\phi(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] \leq \frac{(C-1)K^4}{C^2}(1+\gamma)^2 + \frac{2\epsilon^2 K^2 (C-2)}{C}$$

where $\gamma \triangleq \epsilon\sqrt{\frac{2C}{K^2(C-1)}}$. In particular, as $\epsilon \to 0$, $\gamma \to 0$ and both bounds converge to $\frac{(C-1)K^4}{C^2}$, which depends solely on the inter-class distance K and the number of classes C.

Theorem 4.2 establishes lower and upper bounds on the class-wise distance variance of ID samples under the ETF geometry. Building on the empirical success of Mahalanobis++ in the L2-normalized

feature space and for theoretical tractability, we conduct the subsequent analysis under Assumption 4.1 posed directly on the L2-normalized features $\hat{\phi}$, assuming $\sum_c \boldsymbol{\mu}_c = \mathbf{0}$ so that the ETF structure applies directly to $\boldsymbol{\mu}_c$. Theorem 4.3 further characterizes the conditions under which the class-wise distance variance of L2-normalized ID features strictly exceeds that of OOD samples.

Theorem 4.3 (Sufficient Condition for Strictly Higher Variance in ID Samples)**.** *Let $\hat{\phi}(\mathbf{x}) = (\phi(\mathbf{x}) - \boldsymbol{\mu}_G)/\|\phi(\mathbf{x}) - \boldsymbol{\mu}_G\|_2$ be the centered L2-normalized penultimate feature extractor of a pretrained network. Suppose Assumption 4.1 holds for $\hat{\phi}$ on all $\mathbf{x} \in \mathcal{X}_{in}$, where $R = \|\boldsymbol{\mu}_c\|_2 \leq 1$ and $\sum_c \boldsymbol{\mu}_c = \mathbf{0}$. For any OOD sample $\mathbf{x}^{out} \in \mathcal{X} \setminus \mathcal{X}_{in}$ satisfying $\max_c |\hat{\phi}(\mathbf{x}^{out})^\top \boldsymbol{\mu}_c| < \frac{R(R-\epsilon)}{\sqrt{C-1}}$, we have $\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] > \mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}^{out}) - \boldsymbol{\mu}_c\|_2^2\right]$ for any ID sample $\mathbf{x} \in \mathcal{X}_{in}$.*

Theorem 4.3 shows that when an OOD sample maintains sufficient angular separation from all class means in terms of absolute cosine similarity, the class-wise distance variance of ID samples is guaranteed to strictly exceed that of OOD samples, confirming that the variance term carries meaningful discriminative information for OOD detection. We further strengthen this result in Corollary 4.4 by introducing an additional assumption that the within-class deviation of ID samples lies within the subspace spanned by the class means. Under this assumption, the class-wise distance variance of ID samples is shown to universally dominate that of any OOD sample, regardless of the OOD sample's proximity to the class means.

Corollary 4.4 (Guaranteed Variance Separation under ETF Subspace Alignment)**.** *Under the assumptions of Theorem 4.3, suppose further that for any ID sample $\mathbf{x} \in \mathcal{X}_{in}$ from class c^*, the within-class deviation satisfies $\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_{c^*} \in \mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C)$. Then for any OOD sample $\mathbf{x}^{out} \in \mathcal{X} \setminus \mathcal{X}_{in}$, $\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] \geq \mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}^{out}) - \boldsymbol{\mu}_c\|_2^2\right]$ holds, with equality if and only if $\hat{\phi}(\mathbf{x}^{out}) \in \mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C)$.*

Corollary 4.4 establishes that under the ETF subspace alignment assumption, ID samples universally exhibit higher class-wise distance variance than OOD samples. A notable implication of this result is that even when a near-OOD sample shares the same nearest class distance as an ID sample, the variance term retains discriminative power as long as the near-OOD sample does not lie within the ETF subspace spanned by the class means, a signal that the nearest class distance alone cannot capture. The assumption of Corollary 4.4 corresponds to $\rho(x) = 1$ for ID samples, where $\rho(x) = \|P\hat{\phi}(x)\|_2^2$ and P denotes the projection onto the span of the class means, and it should be regarded as an idealized limit. What we verify empirically is the weaker on-average ordering $\rho_{\mathrm{ID}} > \rho_{\mathrm{OOD}}$, which holds in 22 of 22 backbone and dataset combinations with $D \geq C$ (Appendix D.3). Motivated by the empirical findings in Section 3 and the theoretical analysis presented in this section, we propose **MahaVar**, a post-hoc OOD score that augments the standard Mahalanobis distance with a class-wise distance variance term:

$$S(\mathbf{x}) = -\min_{c \in \{1,\ldots,C\}} d_{\mathrm{Maha},c}(\hat{\phi}(\mathbf{x})) + \alpha \cdot \mathrm{Var}_c\left[d_{\mathrm{Maha},c}(\hat{\phi}(\mathbf{x}))\right] \tag{9}$$

where $\hat{\phi}(\mathbf{x}) = \phi(\mathbf{x})/\|\phi(\mathbf{x})\|_2$ denotes the L2-normalized penultimate feature and $\alpha \geq 0$ is a scalar hyperparameter that controls the contribution of the variance term. Note that while the theoretical analysis is conducted under the ℓ_2 distance for analytical tractability, an analogous bound holds for the Mahalanobis distance under the assumption that the within-class deviation in Mahalanobis sense satisfies $\epsilon_M < K_M/2$, where ϵ_M and K_M denote the Mahalanobis counterparts of ϵ and K, respectively. All proofs are provided in Appendix B. Also, the practical implementation of MahaVar omits the centering step in $\hat{\phi}$; we discuss the rationale and its empirical implications in Appendix D.8.

5 Experiment

Experimental Setup. We follow the benchmark protocol proposed in OpenOOD v1.5. For CIFAR-10 and CIFAR-100 as ID datasets, we use ResNet-18 as the backbone, with MNIST [5], SVHN [30], Textures [3], and Places365 [49] as far-OOD datasets and TIN [19] and CIFAR-100 and CIFAR-10 as near-OOD datasets, respectively. For ImageNet as the ID dataset, we evaluate across three backbone architectures: ResNet-50, Swin-B, and ViT-B, with iNaturalist [41], Textures, and OpenImage-O [43]

Table 1: OOD detection performance on the ImageNet benchmark, averaged over three backbones (ResNet-50, Swin-B, ViT-B). Results are reported in AUROC and FPR@95 for each OOD dataset. Best and second best results are highlighted in **bold** and underline, respectively.

Method	Near-OOD				Far-OOD						Avg	
	SSB-Hard		NINCO		iNaturalist		Texture		OpenImage-O			
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
MSP	70.83	83.15	78.83	73.64	87.68	51.71	82.56	61.60	83.87	61.97	80.75	66.41
MaxLogit	67.47	83.60	74.89	74.46	85.10	52.86	82.30	55.16	80.60	60.71	78.07	65.36
KL-matching	69.38	82.16	80.59	71.49	90.00	46.72	85.14	61.63	87.05	58.53	82.43	64.11
Energy	63.59	85.99	70.01	78.61	79.22	63.69	79.86	56.63	75.35	66.30	73.61	70.24
GEN	72.02	83.48	82.04	71.07	92.46	39.91	87.87	51.96	88.83	53.68	84.65	60.02
ReAct	68.29	84.56	79.14	74.73	90.31	48.05	88.34	49.81	87.31	56.21	82.68	62.67
ASH-S	48.18	92.34	44.57	87.33	40.89	70.66	49.34	69.37	42.95	76.56	45.19	79.25
SCALE	63.33	86.53	69.71	77.14	78.24	56.94	82.04	49.29	77.73	60.58	74.21	66.10
KNN	66.29	88.71	81.03	75.80	88.82	56.87	92.42	36.90	88.23	57.08	83.36	63.07
NCI	67.75	86.91	80.79	74.83	91.50	49.59	89.47	49.00	89.91	54.39	83.88	62.94
NCI+filter	69.56	84.96	81.89	71.74	92.85	40.23	90.67	42.30	90.89	47.95	85.17	57.43
fDBD	68.64	86.32	81.51	73.82	92.02	47.02	90.02	47.55	90.40	53.06	84.52	61.55
ViM	69.63	83.92	82.56	69.67	93.28	39.41	92.43	**34.97**	91.75	45.38	85.93	54.67
NNGuide	71.67	84.11	83.21	70.79	93.48	39.39	92.25	37.71	91.44	48.96	86.41	56.19
Mahalanobis	64.46	85.03	77.99	73.77	84.90	47.86	89.14	51.38	84.70	56.27	80.24	62.86
Mahalanobis++	71.50	78.45	85.56	61.95	95.69	22.22	**93.15**	35.26	92.83	39.24	87.75	47.42
MahaVar (Ours)	**73.02**	**75.87**	**86.92**	**59.03**	**96.82**	**16.12**	93.14	35.83	**93.33**	**36.83**	**88.64**	**44.74**

Table 2: OOD detection performance on CIFAR-10 and CIFAR-100 benchmarks. Results are reported in terms of AUROC and FPR@95, averaged over all OOD datasets for each ID setting. The best and second best results are highlighted in **bold** and underline, respectively.

Method	CIFAR-10		CIFAR-100	
	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓
MSP	89.94	51.34	78.71	79.93
MaxLogit	90.02	40.08	80.25	78.85
KL-matching	87.00	51.70	79.04	77.38
Energy	90.11	**39.78**	80.26	78.26
GEN	90.45	41.39	80.21	78.70
ReAct	89.70	42.95	80.32	78.27
ASH-S	85.60	44.45	81.65	74.96
SCALE	86.61	43.62	81.43	75.71
KNN	**91.81**	45.73	78.35	78.69
NCI	90.38	49.85	81.38	77.84
NCI+filter	90.38	49.85	81.38	77.84
fDBD	88.59	53.37	80.78	77.79
ViM	90.07	57.48	79.16	77.87
NNGuide	90.22	40.37	80.72	78.08
Mahalanobis	88.12	66.50	66.85	90.68
Mahalanobis++	90.80	49.64	82.75	70.33
MahaVar (Ours)	91.12	45.56	**83.80**	**68.03**

as far-OOD datasets and SSB-Hard [42] and NINCO [2] as near-OOD datasets. For the CIFAR-10/100 ID settings, we use the pretrained models provided by OpenOOD v1.5. For the ImageNet ID setting, we use the pretrained models provided by torchvision [27].

We compare MahaVar against a comprehensive set of post-hoc OOD detection baselines: Mahalanobis [20], Mahalanobis++ [29], KNN [39], NCI [23], NCI+filter [23], ViM [43], fDBD [22], MSP [14], MaxLogit [15], KL-matching [15], Energy [24], GEN [25], ReAct [38], ASH-S [6], SCALE [44], and NNGuide [32]. Performance is evaluated using AUROC and FPR@95. All hyperparameters are tuned on the validation sets provided by OpenOOD v1.5, with train, validation, and test splits following its data configuration. The main results for the ImageNet ID setting averaged over three backbone architectures and the CIFAR-10/100 ID settings are presented in Table 1 and Table 2, respectively. Further implementation details are provided in Appendix C.

Main Results. According to Table 1, MahaVar achieves the best AUROC and FPR@95 on all Near-OOD datasets and on two out of three Far-OOD datasets, with Texture being the only exception where MahaVar marginally underperforms Mahalanobis++, confirming that the variance term provides meaningful complementary signal beyond the nearest class distance alone. Notably, while SCALE

and ASH-S achieve the highest performance on ResNet-50, their performance collapses dramatically on Transformer-based backbones, with average AUROC dropping to 63.57% (SCALE) and 23.22% (ASH-S) on Swin-B, and 68.64% (SCALE) and 22.66% (ASH-S) on ViT-B, respectively. In contrast, MahaVar achieves competitive performance on ResNet-50 and attains the best overall performance on Swin-B and ViT-B, consistently outperforming Mahalanobis++ across all backbone architectures and demonstrating stable gains regardless of the underlying feature extractor.

As shown in Table 2, MahaVar consistently outperforms all Mahalanobis-based baselines on both CIFAR settings on average. On CIFAR-100, MahaVar achieves the best overall AUROC and FPR@95 at 83.80% and 68.03%, respectively, outperforming all evaluated baselines. On CIFAR-10, while KNN achieves a higher AUROC, MahaVar outperforms all Mahalanobis-based methods including Mahalanobis++ in both metrics, confirming that the variance term provides reliable gains within the representation-based family. Taken together with the ImageNet results in Table 1, MahaVar achieves state-of-the-art performance on the more challenging ID settings of CIFAR-100 and ImageNet, demonstrating that the variance term provides a reliable complementary signal across diverse benchmark scales. Appendix D provides per-backbone results for the ImageNet benchmark, along with CIFAR-10/100 results against additional baselines.

6 Ablation Study

We conduct ablation studies to validate the design choices of MahaVar, where all ImageNet results are averaged over the three backbone architectures.

Variance term and L2 normalization. Adding the variance term to the standard Mahalanobis distance without L2 normalization already improves the average AUROC from 80.24% to 86.76% and reduces the average FPR@95 from 62.86% to 52.33%, which confirms that the variance term carries meaningful discriminative information. However, this variant still falls short of Mahalanobis++ (87.75% AUROC and 47.42% FPR@95), indicating that L2 normalization stabilizes the feature geometry for distance computation. MahaVar, which combines both components, achieves the best performance overall with 88.64% AUROC and 44.74% FPR@95 (Appendix D.3).

Variance as a standalone score. Using the variance term in isolation yields only 59.91% AUROC and 88.90% FPR@95, indicating that variance alone is insufficient as a standalone score and that the two terms are complementary rather than substitutable. A pointwise analysis further provides a theoretical justification for combining the two terms additively (Appendix D.3).

Higher-order moments. Extending the score with a skewness term yields only marginal gains (88.72% AUROC and 44.56% FPR@95), suggesting that higher-order moments of the class-wise distance distribution contain additional discriminative information. However, since the gains are modest and the optimal sign of the skewness coefficient depends on the backbone, we adopt the variance-only formulation as our main method (Appendix D.4).

Sensitivity to the hyperparameter α. The optimal α is 0.05 for ResNet-50, 0.1 for Swin-B, and 0.07 for ViT-B on the OpenImage-O validation set, and performance degrades as the variance term begins to dominate the nearest-class distance term beyond this range (e.g., the validation AUROC of ResNet-50 drops from 92.75% to 89.30% at $\alpha = 0.2$). This range can be reliably identified using the validation set provided by OpenOOD v1.5, making α selection straightforward in practice (Appendix D.5).

Hyperparameter-free ratio variant. We further explore a ratio-based variant that removes α entirely. On ResNet-50, this score reaches 88.40% AUROC and 39.96% FPR@95 compared with 88.24% and 41.27% for the tuned additive form, though establishing a comparable theoretical separation guarantee for it remains an open direction (Appendix D.9).

Distance metrics. Incorporating the variance term consistently improves performance regardless of the base metric, from 86.67% to 87.50% AUROC for L1, from 87.39% to 88.02% for L2, and from 87.75% to 88.64% for Mahalanobis, confirming the general utility of the variance signal. Among all combinations, the Mahalanobis distance achieves the best results owing to its ability to account for the covariance structure of the feature space (Appendix D.6).

Number of nearest classes. Performance generally improves as the number of nearest classes K used for variance computation increases, with $K = 1000$ being optimal for ResNet-50 and Swin-B, while

ViT-B peaks at $K = 500$ due to its insufficient penultimate layer dimension preventing complete ETF formation. Nevertheless, we adopt $K = 1000$ across all ImageNet backbones in the main experiments for consistency, and all available classes ($K = 10$ and $K = 100$, respectively) are used for CIFAR-10 and CIFAR-100 (Appendix D.7).

When does the variance augmentation help. We provide a practical guideline based on the ratio of feature dimension to class count d/C. The AUROC gain of +1.19 on ResNet-50 ($d/C = 2.05$) and +1.20 on Swin-B ($d/C = 1.02$) shrinks to +0.30 on ViT-B ($d/C = 0.77$), consistent with the incomplete ETF formation observed for ViT-B. Beyond this ratio, we further discuss how the diagnostic statistics of Papyan et al. [31] can directly assess how closely a given ID dataset satisfies the underlying Neural Collapse assumptions (Appendix A).

Comparison with additional recent baselines. Finally, we compare against additional recent baselines not included in the main experiments, including CADRef, LTS, OptFS, and AdaSCALE-A/L (Appendix D.2). On ResNet-50, MahaVar trails activation-shaping and logit-scaling methods such as AdaSCALE, LTS, and CADRef by a narrow margin in AUROC, mirroring the gap observed against SCALE in the main ResNet-50 results. On Swin-B and ViT-B, however, this gap reverses, with MahaVar substantially outperforming all of these methods, consistent with the architecture-dependent trend observed throughout this analysis.

7 Conclusion

We presented MahaVar, a simple and effective post-hoc OOD detector that augments the Mahalanobis distance with a class-wise distance variance term. The key insight is that class-wise distance variance serves as a geometric indicator of class alignment: under Neural Collapse geometry, ID samples concentrate near a specific class mean while remaining far from all others, and this sharp minimum structure directly induces higher class-wise distance variance than OOD samples exhibit. This theoretical grounding, combined with consistent empirical gains across diverse benchmarks and backbone architectures, suggests that the full class-wise distance structure contains richer discriminative information than the nearest-class distance alone can capture. We believe that extending the theoretical framework beyond Neural Collapse, incorporating higher-order moments such as skewness, developing a rigorous treatment of the non-centered feature setting, and establishing a theoretical separation guarantee for ratio-based scoring are promising directions for future work. The limitations of this work are discussed in Appendix E.

Acknowledgments and Disclosure of Funding

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) [grant number RS-2025-02305884] and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) [grant number RS-2026-25619394].

References

[1] Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. NECO: NEural collapse based out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=9ROuKblmi7`.

[2] Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. In *International Conference on Machine Learning*, pages 2471–2506. PMLR, 2023.

[3] 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.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[5] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE signal processing magazine*, 29(6):141–142, 2012.

[6] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ndYXTEL6cZz.

[7] Andrija Djurisic, Rosanne Liu, and Mladen Nikolic. Logit scaling for out-of-distribution detection. *Machine Vision and Applications*, 36(5):108, 2025.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[9] Xuefeng Du, Zhaoning Wang, Mu Cai, and Sharon Li. Vos: Learning what you don't know by virtual outlier synthesis. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TW7d65uYu5M.

[10] Te Han and Yan-Fu Li. Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles. *Reliability Engineering & System Safety*, 226:108648, 2022.

[11] Abid Hassan, Tuan Ngo, Saad Shafiq, and Nenad Medvidovic. Catalyst: Out-of-distribution detection via elastic scaling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pages 1618–1628, June 2026.

[12] Abid Hassan, Tuan Ngo, Saad Shafiq, and Nenad Medvidovic. Davis: Ood detection via dominant activations and variance for increased separation. *arXiv preprint arXiv:2601.22703*, 2026.

[13] 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*, pages 770–778, 2016.

[14] 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. URL https://openreview.net/forum?id=Hkg4TI9xl.

[15] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, pages 8759–8773. PMLR, 2022.

[16] Zesheng Hong, Yubiao Yue, Yubin Chen, Lele Cong, Huanjie Lin, Yuanmei Luo, Mini Han Wang, Weidong Wang, Jialong Xu, Xiaoqi Yang, et al. Out-of-distribution detection in medical image analysis: A survey. *arXiv preprint arXiv:2404.18279*, 2024.

[17] Donghwan Kim, Youngwoong Jin, and Hyunsoo Yoon. Anti-coherence regularized absolute scaling for out-of-distribution detection. *arXiv preprint arXiv:2610.32255*, 2026.

[18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[19] Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[20] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[21] Zhiwei Ling, Yachen Chang, Hailiang Zhao, Xinkui Zhao, Kingsum Chow, and Shuiguang Deng. Cadref: Robust out-of-distribution detection via class-aware decoupled relative feature leveraging. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4968–4977. IEEE, 2025.

[22] Litian Liu and Yao Qin. Fast decision boundary based out-of-distribution detector. In *International Conference on Machine Learning*, pages 31728–31746. PMLR, 2024.

[23] Litian Liu and Yao Qin. Detecting out-of-distribution through the lens of neural collapse. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15424–15433, 2025.

[24] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

[25] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23946–23955, June 2023.

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[27] TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`, 2016.

[28] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=aEFaE0W5pAd`.

[29] Maximilian Müller and Matthias Hein. Mahalanobis++: Improving ood detection via feature normalization. In *International Conference on Machine Learning*, pages 45151–45184. PMLR, 2025.

[30] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

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

[32] 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.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[34] Sandesh Pokhrel, Sanjay Bhandari, Sharib Ali, Tryphon Lambrou, Anh Nguyen, Yash Raj Shrestha, Angus Watson, Danail Stoyanov, Prashnna Gyawali, and Binod Bhattarai. Out-of-distribution detection in gastrointestinal vision by estimating nearest centroid distance deficit. In *Medical Image Understanding and Analysis: 29th Annual Conference, MIUA 2025, Leeds, UK, July 15–17, 2025, Proceedings, Part II*, page 190–200, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-031-98690-1. doi: 10.1007/978-3-031-98691-8_14. URL `https://doi.org/10.1007/978-3-031-98691-8_14`.

[35] Sudarshan Regmi. AdaSCALE: Adaptive scaling for OOD detection. In *Forty-third International Conference on Machine Learning*, 2026. URL `https://openreview.net/forum?id=iXPK0rsCE9`.

[36] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

[37] Youssef Shoeb, Azarm Nowzad, and Hanno Gottschalk. Out-of-distribution segmentation in autonomous driving: Problems and state of the art. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4310–4320, 2025.

[38] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in neural information processing systems*, 34:144–157, 2021.

[39] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International conference on machine learning*, pages 20827–20840. PMLR, 2022.

[40] Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=JHklpEZqduQ`.

[41] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[42] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International conference on learning representations*, 2021.

[43] 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 4921–4930, 2022.

[44] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. Scaling for training time and post-hoc out-of-distribution detection enhancement. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=RDSTjtnqCg`.

[45] Mingyu Xu, Zheng Lian, Bin Liu, and Jianhua Tao. Vra: Variational rectified activation for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 36:28941–28959, 2023.

[46] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

[47] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, et al. Openood v1. 5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2024.

[48] Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. Towards optimal feature-shaping methods for out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=dm8e7gsH0d`.

[49] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[50] Jianing Zhu, Yu Geng, Jiangchao Yao, Tongliang Liu, Gang Niu, Masashi Sugiyama, and Bo Han. Diversified outlier exposure for out-of-distribution detection via informative extrapolation. *Advances in neural information processing systems*, 36:22702–22734, 2023.

A Additional Sorted Class-wise Mahalanobis Distance Structures

This section presents additional sorted class-wise Mahalanobis distance structures and variance distributions to complement the observations in Section 3.2. Specifically, we include results on non-L2-normalized penultimate features for CIFAR-10/100 (Figure 2), and both L2-normalized and non-L2-normalized penultimate features for ImageNet (Figure 3 and Figure 4, respectively). The sharp minimum structure of ID samples is consistently observed across all settings, confirming that the observation in Section 3.2 is not a result of L2 normalization. Additionally, we provide the corresponding $\mathrm{Var}_c[d_{\mathrm{Maha},c}(x)]$ distributions for CIFAR-10/100 without L2 normalization (Figure 5), and for ImageNet with and without L2 normalization (Figure 6 and Figure 7, respectively). Consistent with the main observations, ID samples exhibit higher class-wise distance variance than OOD samples across all settings, with the exception of Transformer-based backbones on ImageNet where the separation is less pronounced due to the insufficient penultimate layer dimension ($d = 768 < C - 1 = 999$ for ViT-B) that prevents complete ETF formation.

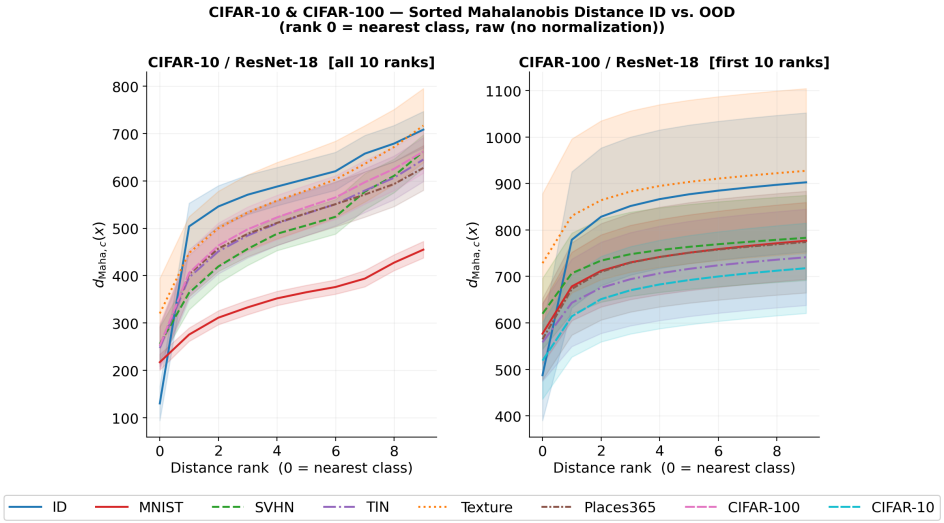


Figure 2: Sorted class-wise $d_{\mathrm{Maha},c}(\mathbf{x})$ (without L2 normalization) on CIFAR-10 and CIFAR-100 as ID datasets, evaluated on OOD datasets following the OpenOOD v1.5 benchmark protocol with ResNet-18 as the backbone. The x-axis represents the Mahalanobis distance rank, where rank 0 corresponds to the nearest class mean. Shaded regions indicate $\pm 0.5\sigma$ across samples at each rank.

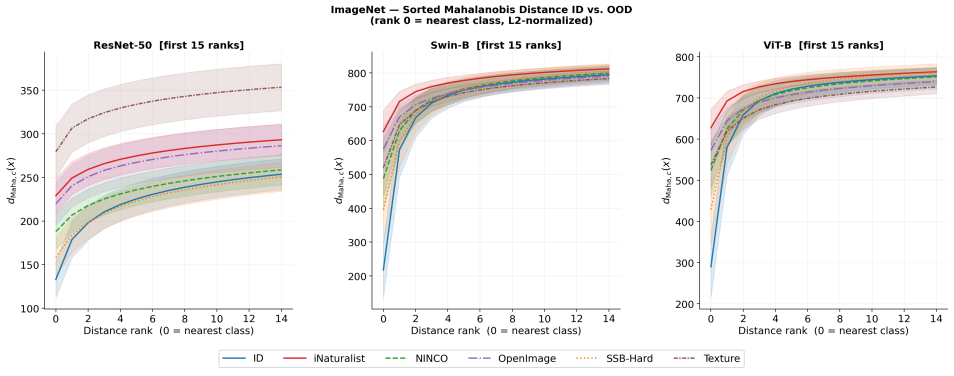


Figure 3: Sorted class-wise $d_{\mathrm{Maha},c}(\mathbf{x})$ of Mahalanobis++ on ImageNet as the ID dataset, evaluated on five OOD datasets following the OpenOOD v1.5 benchmark protocol across three backbone architectures (ResNet-50, Swin-B, and ViT-B). The x-axis represents the Mahalanobis distance rank, where rank 0 corresponds to the nearest class mean. Shaded regions indicate $\pm 0.5\sigma$ across samples at each rank.

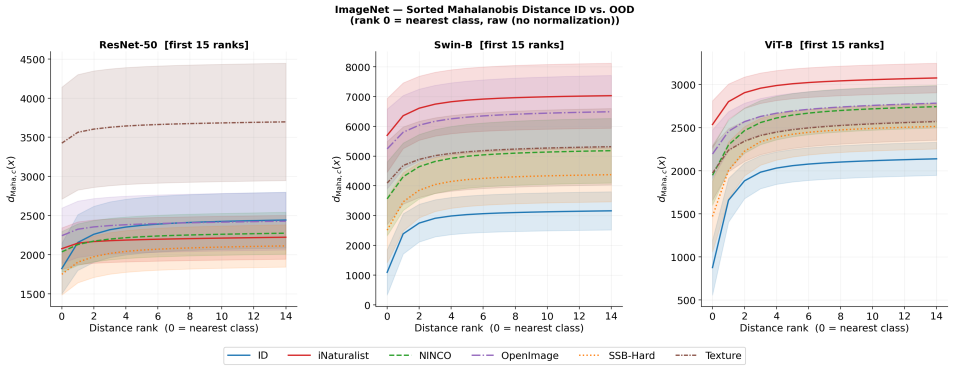


Figure 4: Sorted class-wise $d_{\text{Maha},c}(\mathbf{x})$ (without L2 normalization) on ImageNet as the ID dataset, evaluated on five OOD datasets following the OpenOOD v1.5 benchmark protocol across three backbone architectures (ResNet-50, Swin-B, and ViT-B). The x-axis represents the Mahalanobis distance rank, where rank 0 corresponds to the nearest class mean. Shaded regions indicate $\pm 0.5\sigma$ across samples at each rank.

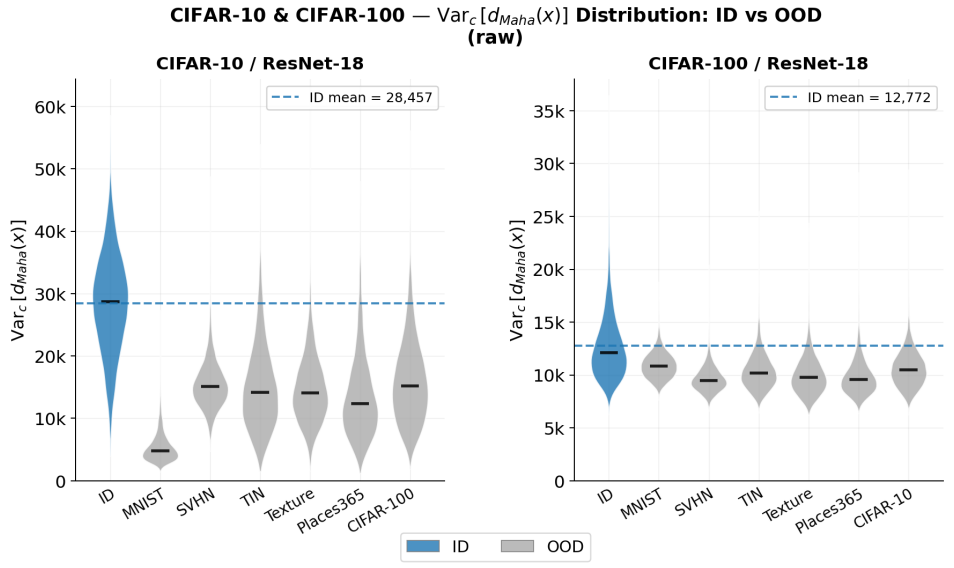


Figure 5: Distribution of $\text{Var}_c[d_{\text{Maha},c}(\mathbf{x})]$ for ID and OOD samples on CIFAR-10 and CIFAR-100 without L2 normalization, evaluated on OOD datasets following the OpenOOD v1.5 benchmark protocol with ResNet-18 as the backbone. The dashed blue line indicates the mean variance of ID samples. Black bars indicate the median of each distribution.

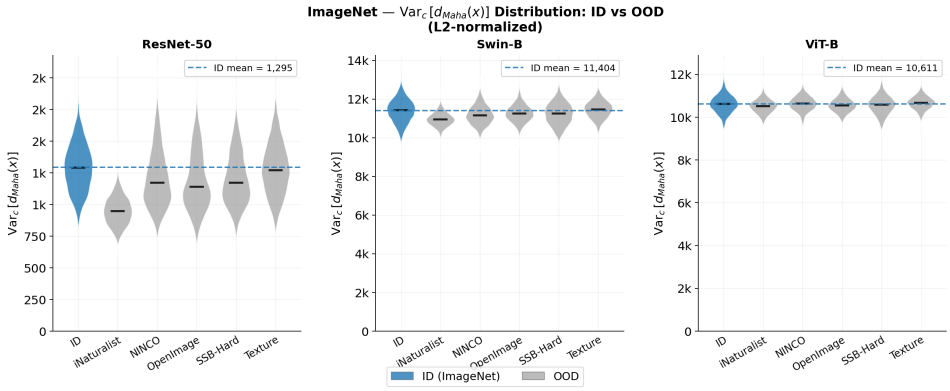


Figure 6: Distribution of $\mathrm{Var}_c[d_{\mathrm{Maha},c}(\mathbf{x})]$ for ID and OOD samples on ImageNet with L2 normalization following Mahalanobis++, evaluated on five OOD datasets following the OpenOOD v1.5 benchmark protocol across three backbone architectures (ResNet-50, Swin-B, and ViT-B). The dashed blue line indicates the mean variance of ID samples. Black bars indicate the median of each distribution.

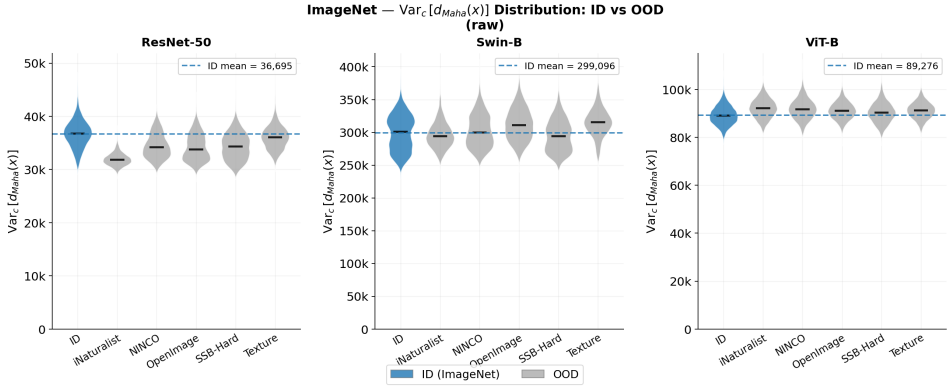


Figure 7: Distribution of $\mathrm{Var}_c[d_{\mathrm{Maha},c}(\mathbf{x})]$ for ID and OOD samples on ImageNet without L2 normalization, evaluated on five OOD datasets following the OpenOOD v1.5 benchmark protocol across three backbone architectures (ResNet-50, Swin-B, and ViT-B). The dashed blue line indicates the mean variance of ID samples. Black bars indicate the median of each distribution.

Additionally, we provide a practical guideline for when the variance augmentation can be expected to help. Table 3 reports the AUROC gain of MahaVar over Mahalanobis++ on ImageNet for each backbone alongside its feature dimension d relative to C.

Table 3: Relationship between the ratio of feature dimension to class count (d/C) and the AUROC gain of MahaVar over Mahalanobis++ on ImageNet, averaged across all five OOD datasets.

Backbone	d	C	d/C	Mahalanobis++ $\rightarrow$ MahaVar (AUROC)
ResNet-50	2048	1000	2.05	87.05 $\rightarrow$ 88.24 $(+1.19)$
Swin-B	1024	1000	1.02	88.29 $\rightarrow$ 89.49 $(+1.20)$
ViT-B	768	1000	0.77	87.90 $\rightarrow$ 88.20 $(+0.30)$

The gain shrinks as d/C drops below 1, consistent with the incomplete ETF formation observed for ViT-B above. This suggests a practical rule of thumb: the variance augmentation is expected to help most when the penultimate feature dimension is at least on the order of the number of classes, that is, $d \gtrsim C$.

Beyond this ratio, one could in principle check for a given ID dataset how closely NC1 and NC2 are satisfied using the diagnostic statistics of Papyan et al. [31]: the within-class variability collapse metric $\mathrm{Tr}(\Sigma_W \Sigma_B^{\dagger})/C$ for NC1, and the coefficient of variation of centered class-mean norms $\mathrm{Std}_c(\|\mu_c - \mu_G\|_2)/\mathrm{Avg}_c(\|\mu_c - \mu_G\|_2)$ together with the standard deviation of pairwise cosine similarities for NC2. However, this guideline can only assess how well the ID data satisfies the underlying assumptions; it says nothing about which OOD samples will be encountered at test time, since these are arbitrary and unknown in advance.

B Proof Theorems

Theorem 4.2 (Bounds on ID Class-wise ℓ_2 Distance Variance)**.** *Let $\phi : \mathcal{X} \to \mathbb{R}^d$ be the penultimate feature extractor of a pretrained network, where $\mathcal{X}_{in} \subset \mathcal{X}$ denotes the in-distribution input space. Suppose Assumption 4.1 holds for ϕ on all $\mathbf{x} \in \mathcal{X}_{in}$. Then for any test ID sample $\mathbf{x} \in \mathcal{X}_{in}$, the class-wise squared ℓ_2 distance variance satisfies:*

$$\frac{(C-1)K^4}{C^2}\left(1-\gamma\right)^2 \leq \mathrm{Var}_c\left[\|\phi(\mathbf{x})-\boldsymbol{\mu}_c\|_2^2\right] \leq \frac{(C-1)K^4}{C^2}\left(1+\gamma\right)^2 + \frac{2\epsilon^2K^2(C-2)}{C}$$

where $\gamma \triangleq \epsilon\sqrt{\frac{2C}{K^2(C-1)}}$. In particular, as $\epsilon \to 0$, $\gamma \to 0$ and both bounds converge to $\frac{(C-1)K^4}{C^2}$, which depends solely on the inter-class distance K and the number of classes C.

Proof. Let $\mathbf{x} \in \mathcal{X}_{in}$ be a test ID sample from class c^*, and denote $\Delta \triangleq \phi(\mathbf{x}) - \boldsymbol{\mu}_{c^*}$ with $\|\Delta\|_2 \leq \epsilon$, and $\delta \triangleq \|\Delta\|_2^2 \leq \epsilon^2$, where the first inequality follows from Assumption 4.1 and the second from $\epsilon = \max_c \epsilon_c$. For any $c \neq c^*$, we write $\phi(\mathbf{x}) - \boldsymbol{\mu}_c = \Delta + (\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_c)$, so that the squared ℓ_2 distance expands as:

$$\|\phi(\mathbf{x})-\boldsymbol{\mu}_c\|_2^2 = \underbrace{\|\Delta\|_2^2}_{=\delta} + \underbrace{\|\boldsymbol{\mu}_{c^*}-\boldsymbol{\mu}_c\|_2^2}_{=K^2 \text{ (by NC2)}} + \underbrace{2\Delta^\top(\boldsymbol{\mu}_{c^*}-\boldsymbol{\mu}_c)}_{\triangleq \xi_c} \tag{10}$$

Letting $\bar{\xi} \triangleq \frac{1}{C-1}\sum_{c\neq c^*}\xi_c$, the mean class-wise squared distance becomes:

$$\bar{d} = \frac{1}{C}\left[\delta + \sum_{c\neq c^*}(\delta + K^2 + \xi_c)\right] = \delta + \frac{(C-1)(K^2+\bar{\xi})}{C} \tag{11}$$

where the first term δ is the contribution from $c = c^*$, and the second term averages the remaining $C-1$ distances. The deviation from $\bar{d}$ for $c = c^*$ is then:

$$\|\phi(\mathbf{x})-\boldsymbol{\mu}_{c^*}\|_2^2 - \bar{d} = \delta - \delta - \frac{(C-1)(K^2+\bar{\xi})}{C} = -\frac{(C-1)(K^2+\bar{\xi})}{C} \tag{12}$$

and for $c \neq c^*$:

$$\begin{aligned}
\|\phi(\mathbf{x})-\boldsymbol{\mu}_c\|_2^2 - \bar{d} &= (\delta + K^2 + \xi_c) - \delta - \frac{(C-1)(K^2+\bar{\xi})}{C} \\
&= K^2 + \xi_c - \frac{(C-1)(K^2+\bar{\xi})}{C} \\
&= \frac{C(K^2+\xi_c) - (C-1)(K^2+\bar{\xi})}{C} \\
&= \frac{K^2+\bar{\xi}}{C} + (\xi_c - \bar{\xi})
\end{aligned} \tag{13}$$

Letting $S_\xi^2 \triangleq \frac{1}{C-1}\sum_{c\neq c^*}(\xi_c-\bar{\xi})^2 \geq 0$, summing squared deviations over all C classes and using $\sum_{c\neq c^*}(\xi_c - \bar{\xi}) = 0$ by definition of $\bar{\xi}$:

$$\begin{aligned}
C \cdot \mathrm{Var}_c\left[\|\phi(\mathbf{x})-\boldsymbol{\mu}_c\|_2^2\right] &= \frac{(C-1)^2(K^2+\bar{\xi})^2}{C^2} + (C-1)\left(\frac{K^2+\bar{\xi}}{C}\right)^2 + (C-1)S_\xi^2 \\
&= \frac{(C-1)(K^2+\bar{\xi})^2}{C} + (C-1)S_\xi^2
\end{aligned} \tag{14}$$

Dividing both sides by C, the variance reduces to:

$$\mathrm{Var}_c\left[\|\phi(\mathbf{x})-\boldsymbol{\mu}_c\|_2^2\right] = \frac{(C-1)(K^2+\bar{\xi})^2}{C^2} + \frac{(C-1)S_\xi^2}{C} \tag{15}$$

We now derive tighter bounds on $\bar{\xi}$ and S_ξ^2 by fully exploiting the simplex ETF structure. Since $\boldsymbol{\mu}_G = \frac{1}{C}\sum_c \boldsymbol{\mu}_c$, we have $\sum_{c\neq c^*} \boldsymbol{\mu}_c = C\boldsymbol{\mu}_G - \boldsymbol{\mu}_{c^*}$, so that:

$$\sum_{c\neq c^*} (\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_c) = (C-1)\boldsymbol{\mu}_{c^*} - (C\boldsymbol{\mu}_G - \boldsymbol{\mu}_{c^*}) = C(\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_G) \tag{16}$$

Substituting this into the definition of $\bar{\xi}$ and using linearity of the inner product:

$$\begin{aligned}\bar{\xi} &= \frac{1}{C-1}\sum_{c\neq c^*} \xi_c = \frac{2}{C-1}\Delta^\top \sum_{c\neq c^*} (\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_c) \\ &= \frac{2}{C-1}\Delta^\top \cdot C(\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_G) = \frac{2C}{C-1}\Delta^\top(\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_G)\end{aligned} \tag{17}$$

Denoting $R \triangleq \|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2$ for all c (equinorm by NC2), we first express R in terms of K. By the simplex ETF condition in Assumption 4.1, the centered class means satisfy:

$$\frac{(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)^\top(\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G)}{\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2\|\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G\|_2} = -\frac{1}{C-1} \quad \forall c \neq c' \tag{18}$$

Since $\|\boldsymbol{\mu}_c - \boldsymbol{\mu}_G\|_2 = R$ for all c, this gives:

$$(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)^\top(\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G) = -\frac{R^2}{C-1} \quad \forall c \neq c' \tag{19}$$

Substituting into the ETF pairwise distance:

$$\begin{aligned}K^2 = \|\boldsymbol{\mu}_c - \boldsymbol{\mu}_{c'}\|_2^2 &= \|(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G) - (\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G)\|_2^2 \\ &= 2R^2 - 2(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)^\top(\boldsymbol{\mu}_{c'} - \boldsymbol{\mu}_G) \\ &= 2R^2 - 2\left(-\frac{R^2}{C-1}\right) = 2R^2 \cdot \frac{C}{C-1}\end{aligned} \tag{20}$$

which gives $R^2 = \frac{(C-1)K^2}{2C}$. Applying Cauchy-Schwarz to the exact expression of $\bar{\xi}$:

$$|\bar{\xi}| \leq \frac{2C}{C-1}\|\Delta\|_2 \cdot R \leq \frac{2C\epsilon}{C-1} \cdot K\sqrt{\frac{C-1}{2C}} = \epsilon K\sqrt{\frac{2C}{C-1}} \tag{21}$$

For S_ξ^2, we derive a bound by expanding $\xi_c - \bar{\xi}$. Recall that $\xi_c = 2\Delta^\top(\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_c)$ and $\bar{\xi} = \frac{2C}{C-1}\Delta^\top(\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_G)$. Hence, we can derive following:

$$\begin{aligned}\xi_c - \bar{\xi} &= 2\Delta^\top(\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_c) - \frac{2C}{C-1}\Delta^\top(\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_G) \\ &= 2\Delta^\top\left[(\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_c) - \frac{C}{C-1}(\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_G)\right] \\ &= 2\Delta^\top\left[-\frac{\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_G}{C-1} - (\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)\right] \\ &= -2\Delta^\top \mathbf{v}_c\end{aligned} \tag{22}$$

where $\mathbf{v}_c \triangleq \frac{\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_G}{C-1} + (\boldsymbol{\mu}_c - \boldsymbol{\mu}_G)$.

Using the ETF inner product $(\boldsymbol{\mu}_{c^*} - \boldsymbol{\mu}_G)^\top(\boldsymbol{\mu}_c - \boldsymbol{\mu}_G) = -\frac{R^2}{C-1}$ for $c \neq c^*$:

$$\begin{aligned}\|\mathbf{v}_c\|_2^2 &= \frac{R^2}{(C-1)^2} - \frac{2R^2}{(C-1)^2} + R^2 \\ &= R^2\left(1 - \frac{1}{(C-1)^2}\right) \\ &= \frac{R^2C(C-2)}{(C-1)^2} = \frac{K^2(C-2)}{2(C-1)}\end{aligned} \tag{23}$$

Applying Cauchy-Schwarz to each term $(\xi_c - \bar{\xi})^2 = (-2\Delta^\top \mathbf{v}_c)^2 \le 4\|\Delta\|_2^2\|\mathbf{v}_c\|_2^2 \le 4\epsilon^2\|\mathbf{v}_c\|_2^2$ and summing over $c \neq c^*$:

$$S_\xi^2 = \frac{1}{C-1}\sum_{c\neq c^*}(\xi_c - \bar{\xi})^2 \le \frac{4\epsilon^2}{C-1}\sum_{c\neq c^*}\|\mathbf{v}_c\|_2^2 = 4\epsilon^2 \cdot \frac{K^2(C-2)}{2(C-1)} = \frac{2\epsilon^2K^2(C-2)}{C-1}. \tag{24}$$

To derive the lower bound, we substitute $S_\xi^2 \ge 0$ and $\bar{\xi} \ge -\epsilon K\sqrt{\frac{2C}{C-1}}$ into Eq. (15):

$$\begin{aligned}\mathrm{Var}_c\left[\|\phi(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] &= \frac{(C-1)(K^2+\bar{\xi})^2}{C^2} + \frac{(C-1)S_\xi^2}{C} \\ &\ge \frac{(C-1)(K^2+\bar{\xi})^2}{C^2} \\ &\ge \frac{(C-1)\left(K^2 - \epsilon K\sqrt{\frac{2C}{C-1}}\right)^2}{C^2} \\ &= \frac{(C-1)K^4\left(1 - \frac{\epsilon}{K}\sqrt{\frac{2C}{C-1}}\right)^2}{C^2} \\ &= \frac{(C-1)K^4}{C^2}\left(1 - \epsilon\sqrt{\frac{2C}{K^2(C-1)}}\right)^2\end{aligned} \tag{25}$$

where the third inequality holds since $K^2 + \bar{\xi} > 0$ for all admissible $\bar{\xi}$, as $\epsilon < K/2$ implies $K^2 - \epsilon K\sqrt{\frac{2C}{C-1}} \ge K\left(K - \epsilon \cdot 2\right) > 0$, so $(K^2+\bar{\xi})^2$ is minimized at $\bar{\xi} = -\epsilon K\sqrt{\frac{2C}{C-1}}$. For the upper bound, substituting $\bar{\xi} \le \epsilon K\sqrt{\frac{2C}{C-1}}$ and $S_\xi^2 \le \frac{2\epsilon^2K^2(C-2)}{C-1}$ into Eq. (15):

$$\mathrm{Var}_c\left[\|\phi(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] \le \frac{(C-1)K^4}{C^2}\left(1 + \epsilon\sqrt{\frac{2C}{K^2(C-1)}}\right)^2 + \frac{2\epsilon^2K^2(C-2)}{C} \tag{26}$$

Finally, as $\epsilon \to 0$, all terms involving ϵ vanish, and Eq. (15) converges to $\frac{(C-1)K^4}{C^2}$, completing the proof. □

Theorem 4.3 (Sufficient Condition for Strictly Higher Variance in ID Samples)**.** *Let $\hat{\phi}(\mathbf{x}) = (\phi(\mathbf{x}) - \boldsymbol{\mu}_G)/\|\phi(\mathbf{x}) - \boldsymbol{\mu}_G\|_2$ be the centered L2-normalized penultimate feature extractor of a pretrained network. Suppose Assumption 4.1 holds for $\hat{\phi}$ on all $\mathbf{x} \in \mathcal{X}_{in}$, where $R = \|\boldsymbol{\mu}_c\|_2 \le 1$ and $\sum_c \boldsymbol{\mu}_c = \mathbf{0}$. For any OOD sample $\mathbf{x}^{out} \in \mathcal{X} \setminus \mathcal{X}_{in}$ satisfying $\max_c |\hat{\phi}(\mathbf{x}^{out})^\top \boldsymbol{\mu}_c| < \frac{R(R-\epsilon)}{\sqrt{C-1}}$, we have $\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] > \mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}^{out}) - \boldsymbol{\mu}_c\|_2^2\right]$ for any ID sample $\mathbf{x} \in \mathcal{X}_{in}$.*

Proof. We establish the result by deriving a lower bound on $\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right]$ for ID samples and an upper bound on $\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}^{out}) - \boldsymbol{\mu}_c\|_2^2\right]$ for OOD samples. Since $\|\boldsymbol{\mu}_c\|_2 = R$ for all c, the

ETF inter-class distance satisfies $K^2 = \frac{2R^2C}{C-1}$. Substituting into Theorem 4.2 gives $\frac{(C-1)K^4}{C^2} = \frac{4R^4}{C-1}$ and $\gamma = \frac{\epsilon}{R}$, so that:

$$\begin{aligned}\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] &\geq \frac{(C-1)K^4}{C^2}(1-\gamma)^2 \\ &= \frac{4R^4}{C-1}\left(1-\frac{\epsilon}{R}\right)^2 \\ &= \frac{4R^4}{C-1}\cdot\frac{(R-\epsilon)^2}{R^2} \\ &= \frac{4R^2(R-\epsilon)^2}{C-1}\end{aligned} \tag{27}$$

For the OOD upper bound, since $\|\hat{\phi}(\mathbf{x}^{out})\|_2 = 1$ and $\|\boldsymbol{\mu}_c\|_2 = R$ for all c, we have $\|\hat{\phi}(\mathbf{x}^{out}) - \boldsymbol{\mu}_c\|_2^2 = 1 + R^2 - 2s_c$ where $s_c \triangleq \hat{\phi}(\mathbf{x}^{out})^\top \boldsymbol{\mu}_c$. Since R^2 is constant across classes, it vanishes under the variance operator, giving:

$$\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}^{out}) - \boldsymbol{\mu}_c\|_2^2\right] = 4\,\mathrm{Var}_c\left[s_c\right] \tag{28}$$

From $\sum_c \boldsymbol{\mu}_c = \mathbf{0}$ it follows that $\sum_c s_c = \hat{\phi}(\mathbf{x}^{out})^\top \sum_c \boldsymbol{\mu}_c = 0$ and $\mathrm{Var}_c[s_c] = \frac{1}{C}\sum_c s_c^2$. To bound $\mathrm{Var}_c[s_c]$, we introduce $t \triangleq \max_c |s_c|$ as the worst-case absolute cosine similarity between the OOD sample and any class mean. Since $|s_c| \leq t$ for all c, we have $s_c^2 \leq t^2$ for all c, and therefore $\mathrm{Var}_c[s_c] \leq \frac{1}{C}\sum_c t^2 = t^2$, giving:

$$\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}^{out}) - \boldsymbol{\mu}_c\|_2^2\right] \leq 4t^2 \tag{29}$$

The ID lower bound exceeds the OOD upper bound when:

$$\frac{4R^2(R-\epsilon)^2}{C-1} > 4t^2 \iff t < \frac{R(R-\epsilon)}{\sqrt{C-1}} \tag{30}$$

which holds under the stated condition $\max_c |s_c| < \frac{R(R-\epsilon)}{\sqrt{C-1}}$, completing the proof. □

Corollary 4.4 (Guaranteed Variance Separation under ETF Subspace Alignment)**.** *Under the assumptions of Theorem 4.3, suppose further that for any ID sample $\mathbf{x} \in \mathcal{X}_{in}$ from class c^*, the within-class deviation satisfies $\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_{c^*} \in \mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C)$. Then for any OOD sample $\mathbf{x}^{out} \in \mathcal{X} \setminus \mathcal{X}_{in}$, $\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] \geq \mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}^{out}) - \boldsymbol{\mu}_c\|_2^2\right]$ holds, with equality if and only if $\hat{\phi}(\mathbf{x}^{out}) \in \mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C)$.*

Proof. We first establish the exact expression $\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] = \frac{4R^2\rho}{C-1}$ where $\rho \triangleq \|P\hat{\phi}(\mathbf{x})\|_2^2$ and P denotes the orthogonal projection onto $\mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C)$. Let $M = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C]^\top \in \mathbb{R}^{C\times d}$ denote the matrix of class means stacked as rows, and let $\mathbf{1} \in \mathbb{R}^C$ denote the all-ones vector. Since $\|\hat{\phi}(\mathbf{x})\|_2 = 1$ and $\|\boldsymbol{\mu}_c\|_2 = R$ for all c, we have $d_c^2 = \|\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2 = 1 + R^2 - 2s_c$ where $s_c \triangleq \hat{\phi}(\mathbf{x})^\top \boldsymbol{\mu}_c$. Since R^2 is constant across classes, it vanishes under the variance operator, giving $\mathrm{Var}_c[d_c^2] = 4\mathrm{Var}_c[s_c]$. From $\sum_c \boldsymbol{\mu}_c = \mathbf{0}$:

$$\sum_c s_c = \hat{\phi}(\mathbf{x})^\top \sum_c \boldsymbol{\mu}_c = 0 \implies \mathrm{Var}_c[s_c] = \frac{1}{C}\sum_c s_c^2 = \frac{1}{C}\hat{\phi}(\mathbf{x})^\top M^\top M \hat{\phi}(\mathbf{x}) \tag{31}$$

We now derive $M^\top M = \frac{R^2 C}{C-1} P$. Computing $MM^\top \in \mathbb{R}^{C\times C}$ using the ETF conditions $(MM^\top)_{c,c} = R^2$ and $(MM^\top)_{c,c'} = -\frac{R^2}{C-1}$ for $c \neq c'$:

$$\begin{aligned} MM^\top &= R^2 I_C + \left(-\frac{R^2}{C-1}\right)(\mathbf{1}\mathbf{1}^\top - I_C) \\ &= R^2 I_C - \frac{R^2}{C-1}\mathbf{1}\mathbf{1}^\top + \frac{R^2}{C-1} I_C \\ &= R^2\left(1 + \frac{1}{C-1}\right) I_C - \frac{R^2}{C-1}\mathbf{1}\mathbf{1}^\top \\ &= \frac{R^2 C}{C-1} I_C - \frac{R^2}{C-1}\mathbf{1}\mathbf{1}^\top \\ &= \frac{R^2 C}{C-1}\left(I_C - \frac{1}{C}\mathbf{1}\mathbf{1}^\top\right) \\ &= \frac{R^2 C}{C-1} H \end{aligned} \tag{32}$$

where $H = I_C - \frac{1}{C}\mathbf{1}\mathbf{1}^\top$ is the centering matrix. We verify that H is idempotent using $(\mathbf{1}\mathbf{1}^\top)^2 = C\mathbf{1}\mathbf{1}^\top$:

$$H^2 = I_C - \frac{2}{C}\mathbf{1}\mathbf{1}^\top + \frac{1}{C^2}(\mathbf{1}\mathbf{1}^\top)^2 = I_C - \frac{2}{C}\mathbf{1}\mathbf{1}^\top + \frac{1}{C^2}\cdot C\mathbf{1}\mathbf{1}^\top = I_C - \frac{1}{C}\mathbf{1}\mathbf{1}^\top = H \tag{33}$$

Since H is idempotent and symmetric, $(MM^\top)^+ = \frac{C-1}{R^2 C} H$, so that:

$$P = M^\top (MM^\top)^+ M = \frac{C-1}{R^2 C} M^\top H M = \frac{C-1}{R^2 C}\left(M^\top M - \frac{1}{C} M^\top \mathbf{1}\mathbf{1}^\top M\right) \tag{34}$$

Since $M^\top \mathbf{1} = \sum_c \boldsymbol{\mu}_c = \mathbf{0}$, we have $M^\top \mathbf{1}\mathbf{1}^\top M = \mathbf{0}$, and therefore:

$$P = \frac{C-1}{R^2 C} M^\top M \implies M^\top M = \frac{R^2 C}{C-1} P \tag{35}$$

Plugging this into the expression for $\mathrm{Var}_c[s_c]$:

$$\mathrm{Var}_c[s_c] = \frac{1}{C}\hat{\phi}(\mathbf{x})^\top \cdot \frac{R^2 C}{C-1} P \cdot \hat{\phi}(\mathbf{x}) = \frac{R^2 \rho}{C-1} \implies \mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] = \frac{4R^2\rho}{C-1} \tag{36}$$

For any ID sample $\mathbf{x} \in \mathcal{X}_{in}$, the within-class deviation assumption gives $\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_{c^*} \in \mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C)$. Since $\boldsymbol{\mu}_{c^*}$ trivially lies in $\mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C)$, it follows that $\hat{\phi}(\mathbf{x}) \in \mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C)$, hence $P\hat{\phi}(\mathbf{x}) = \hat{\phi}(\mathbf{x})$ and:

$$\rho_{ID} = \|P\hat{\phi}(\mathbf{x})\|_2^2 = \|\hat{\phi}(\mathbf{x})\|_2^2 = 1 \implies \mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] = \frac{4R^2}{C-1} \tag{37}$$

For any OOD sample $\mathbf{x}^{out} \in \mathcal{X} \setminus \mathcal{X}_{in}$, since P is an orthogonal projection and $\hat{\phi}(\mathbf{x}^{out})$ lies on the unit sphere, $\rho_{OOD} = \hat{\phi}(\mathbf{x}^{out})^\top P \hat{\phi}(\mathbf{x}^{out}) \leq \|\hat{\phi}(\mathbf{x}^{out})\|_2^2 = 1$, and therefore:

$$\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}^{out}) - \boldsymbol{\mu}_c\|_2^2\right] = \frac{4R^2\rho_{OOD}}{C-1} \leq \frac{4R^2}{C-1} \tag{38}$$

Hence $\mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}) - \boldsymbol{\mu}_c\|_2^2\right] \geq \mathrm{Var}_c\left[\|\hat{\phi}(\mathbf{x}^{out}) - \boldsymbol{\mu}_c\|_2^2\right]$ holds, with equality if and only if $\rho_{OOD} = 1$, i.e., $\hat{\phi}(\mathbf{x}^{out}) \in \mathrm{span}(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C)$, completing the proof. □

C Implementation Details

This section provides implementation details and hyperparameter configurations for all methods evaluated in the main experiments. All experiments are implemented in PyTorch 2.4.1 [33].

Mahalanobis and Mahalanobis++. Class means $\hat{\boldsymbol{\mu}}_c$ and a shared tied covariance matrix $\hat{\Sigma}$ are estimated from the penultimate features of the training set. To prevent ill-conditioning, a regularization term $10^{-3} \cdot \mathbf{I}$ is added to the estimated covariance matrix. The OOD score is computed using the inverse of the covariance matrix:

$$S(\mathbf{x}) = -\min_c d_{\mathrm{Maha},c}(\mathbf{x}), \quad d_{\mathrm{Maha},c}(\mathbf{x}) = (\hat{\phi}(\mathbf{x}) - \hat{\boldsymbol{\mu}}_c)^\top \hat{\Sigma}^{-1} (\hat{\phi}(\mathbf{x}) - \hat{\boldsymbol{\mu}}_c), \tag{39}$$

where $\phi(\mathbf{x}) \in \mathbb{R}^d$ denotes the penultimate layer feature of input $\mathbf{x}$, and the terms $\hat{\Sigma}^{-1}\hat{\boldsymbol{\mu}}_c$ and $\hat{\boldsymbol{\mu}}_c^\top \hat{\Sigma}^{-1} \hat{\boldsymbol{\mu}}_c$ are precomputed for inference efficiency. For Mahalanobis++, features are L2-normalized prior to all computations: $\hat{\phi}(\mathbf{x}) = \phi(\mathbf{x})/\|\phi(\mathbf{x})\|_2$.

MahaVar. MahaVar follows the same preprocessing and class statistics as Mahalanobis++, with the OOD score augmented by a class-wise distance variance term:

$$S(\mathbf{x}) = -\min_c d_{\mathrm{Maha},c}(\hat{\phi}(\mathbf{x})) + \alpha \cdot \mathrm{Var}_c\big[d_{\mathrm{Maha},c}(\hat{\phi}(\mathbf{x}))\big]. \tag{40}$$

The hyperparameter $\alpha \geq 0$ controls the relative contribution of the variance term. α is selected via grid search over 26 values in $[0, 10]$ using the validation AUROC provided by OpenOOD v1.5: OpenImage-O for ImageNet settings and TIN for CIFAR settings. The selected values are $\alpha = 0.05$ (ResNet-50), 0.1 (Swin-B), 0.07 (ViT-B), 0.015 (CIFAR-10), and 0.03 (CIFAR-100). Since the class means $\hat{\boldsymbol{\mu}}_c$ and inverse covariance matrix $\hat{\Sigma}^{-1}$ are precomputed and stored, the class-wise distance variance at inference time is computed directly from the per-sample Mahalanobis distances with negligible additional overhead. In practice, we observe that MahaVar achieves inference latency identical to that of Mahalanobis++.

MSP. MSP [14] uses the maximum class probability after softmax as the OOD score:

$$S(\mathbf{x}) = \max_c \mathrm{softmax}(\mathbf{W}\phi(\mathbf{x}) + \mathbf{b})_c, \tag{41}$$

where $\mathbf{W} \in \mathbb{R}^{C \times d}$ and $\mathbf{b} \in \mathbb{R}^C$ denote the classifier head weight matrix and bias vector, respectively.

MaxLogit. MaxLogit [15] replaces the softmax with the maximum raw logit:

$$S(\mathbf{x}) = \max_c (\mathbf{W}\phi(\mathbf{x}) + \mathbf{b})_c. \tag{42}$$

KL-Matching. KL-Matching [15] computes class-conditional reference distributions $\mathbf{d}_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathrm{softmax}(\mathbf{W}\phi(\mathbf{x}_i) + \mathbf{b})$ from the training set, and scores a test sample by:

$$S(\mathbf{x}) = -\min_c \mathrm{KL}(\mathbf{p}(\mathbf{x}) \,\|\, \mathbf{d}_c), \tag{43}$$

where $\mathbf{p}(\mathbf{x}) = \mathrm{softmax}(\mathbf{W}\phi(\mathbf{x}) + \mathbf{b})$ is the predicted class probability vector.

Energy. The energy score [24] aggregates all class logits via log-sum-exp:

$$S(\mathbf{x}) = T \cdot \log \sum_c \exp\left(\frac{(\mathbf{W}\phi(\mathbf{x}) + \mathbf{b})_c}{T}\right), \tag{44}$$

with temperature $T = 1$ fixed throughout.

GEN. GEN [25] computes a generalized entropy over the top-M class probabilities:

$$S(\mathbf{x}) = -\sum_{m \in \text{top-}M} p_m^\gamma (1 - p_m)^\gamma, \tag{45}$$

where p_m denotes the m-th largest entry of $\mathbf{p}(\mathbf{x})$, and γ controls the sharpness of the entropy measure. $M = 100$ is fixed for all settings except CIFAR-10, where $M = 10$. γ is selected from $\{0.05, 0.10, 0.20, 0.30, 0.50, 0.70, 1.00\}$ via validation AUROC; the selected values are $\gamma = 0.05$ (ResNet-50), 0.5 (Swin-B), 0.3 (ViT-B), 0.2 (CIFAR-10), and 0.5 (CIFAR-100).

ReAct. ReAct [38] clips penultimate activations at a threshold τ set to the p-th percentile of the training features, prior to energy scoring:

$$S(\mathbf{x}) = T \cdot \log \sum_c \exp\left(\frac{(\mathbf{W}\,\mathrm{clip}(\phi(\mathbf{x}), \tau) + \mathbf{b})_c}{T}\right), \tag{46}$$

where $\mathrm{clip}(\phi(\mathbf{x}), \tau)$ truncates each activation at τ from above. The percentile p is selected from $\{0.90, 0.95, 0.99\}$ via validation AUROC; the selected values are $p = 0.95$ / $\tau = 1.351$ (ResNet-50), $p = 0.90$ / $\tau = 0.376$ (Swin-B), $p = 0.90$ / $\tau = 0.692$ (ViT-B), $p = 0.99$ / $\tau = 1.056$ (CIFAR-10), and $p = 0.99$ / $\tau = 2.278$ (CIFAR-100).

ASH-S and SCALE. ASH-S [6] prunes activations below a threshold τ and applies exponential rescaling:

$$\phi_{\mathrm{ASH}}(\mathbf{x}) = \phi(\mathbf{x}) \cdot \mathbf{1}[\phi(\mathbf{x}) > \tau] \cdot \exp\left(\frac{\sum_i \phi_i(\mathbf{x})}{\sum_i \phi_i(\mathbf{x})\mathbf{1}[\phi_i(\mathbf{x}) > \tau]}\right), \tag{47}$$

followed by energy scoring. SCALE [44] applies the same exponential rescaling without pruning:

$$\phi_{\mathrm{SCALE}}(\mathbf{x}) = \phi(\mathbf{x}) \cdot \exp\left(\frac{\sum_i \phi_i(\mathbf{x})}{\sum_i \phi_i(\mathbf{x})\mathbf{1}[\phi_i(\mathbf{x}) > \tau]}\right), \tag{48}$$

followed by energy scoring. For both methods, the pruning threshold τ is determined by the p-th percentile of each test sample's own activations, with p selected from $\{0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$ via validation AUROC. For ASH-S, the selected values are $p = 0.85$ (ResNet-50), 0.65 (Swin-B), 0.65 (ViT-B), 0.65 (CIFAR-10), and 0.65 (CIFAR-100). For SCALE, the selected values are $p = 0.85$ (ResNet-50), 0.95 (Swin-B), 0.65 (ViT-B), 0.65 (CIFAR-10), and 0.65 (CIFAR-100).

KNN. KNN [39] scores a test sample by its cosine distance to the k-th nearest neighbor in a subsampled training index:

$$S(\mathbf{x}) = -(1 - \hat{\phi}(\mathbf{x})^\top \hat{\mathbf{z}}_k), \tag{49}$$

where $\hat{\mathbf{z}}_k \in \mathbb{R}^d$ is the k-th nearest L2-normalized training feature. The index is built from a random 1% subsample of the training set across all settings. k is selected from $\{1, 3, 5, 10, 20, 50\}$ via validation AUROC; the selected value is $k = 3$ across all ImageNet backbones. For CIFAR-10, $k = 3$, and for CIFAR-100, $k = 1$.

NCI and NCI+filter. NCI [23] scores a test sample by the alignment between its centered feature and the predicted class weight vector:

$$S(\mathbf{x}) = \frac{(\phi(\mathbf{x}) - \hat{\boldsymbol{\mu}}_G)^\top}{\|\phi(\mathbf{x}) - \hat{\boldsymbol{\mu}}_G\|_2}\mathbf{w}_{\hat{c}} + \alpha\|\phi(\mathbf{x})\|_1, \tag{50}$$

where $\hat{c} = \arg\max_c(\mathbf{W}\phi(\mathbf{x}) + \mathbf{b})_c$ is the predicted class, $\hat{\boldsymbol{\mu}}_G = \frac{1}{N}\sum_i \phi(\mathbf{x}_i)$ is the global training mean, and $\mathbf{w}_{\hat{c}} \in \mathbb{R}^d$ is the weight vector of the predicted class. The weight α on the L1 norm filter term is selected from the same grid as MahaVar via validation AUROC; the selected values are $\alpha = 5 \times 10^{-4}$ (ResNet-50), 1×10^{-4} (Swin-B), 3×10^{-4} (ViT-B), and 0.0 for both CIFAR settings. NCI+filter additionally applies a feature norm threshold selected on the validation set.

fDBD. fDBD [22] scores a test sample by its normalized distance to the decision boundaries between the predicted class and all other classes:

$$S(\mathbf{x}) = \frac{1}{\|\phi(\mathbf{x}) - \hat{\boldsymbol{\mu}}_G\|_2} \cdot \frac{1}{C-1}\sum_{c \neq \hat{c}} \frac{|l_{\hat{c}}(\mathbf{x}) - l_c(\mathbf{x})|}{\|\mathbf{w}_{\hat{c}} - \mathbf{w}_c\|_2} \tag{51}$$

where $l_c(\mathbf{x}) = \mathbf{w}_c^\top \phi(\mathbf{x}) + b_c$ denotes the logit for class c. The pairwise weight distances $\|\mathbf{w}_{\hat{c}} - \mathbf{w}_c\|_2$ are precomputed as a $C \times C$ matrix. fDBD requires no hyperparameter tuning.

ViM. ViM [43] constructs a virtual OOD logit from the residual of the feature against the top-D principal subspace of the training set, computed via randomized SVD on a subsample of 200,000 training features. The virtual logit is scaled to be commensurate with the class logits and appended to the logit vector prior to softmax scoring. The principal subspace dimension D is selected from $\{64, 128, 256\}$ via validation AUROC; the selected value is $D = 256$ across all ImageNet backbones, $D = 128$ for CIFAR-10, and $D = 64$ for CIFAR-100.

NNGuide. NNGuide [32] guides a base confidence score by the average confidence-weighted cosine similarity to the k-nearest neighbors in a bank set:

$$S(\mathbf{x}) = S_{\text{base}}(\mathbf{x}) \cdot G(\mathbf{x}), \quad G(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{k} s_{(i)} \, \text{sim}(\mathbf{z}_{(i)}, \mathbf{z}), \tag{52}$$

where $\mathbf{z} = \phi(\mathbf{x})/\|\phi(\mathbf{x})\|_2$ is the L2-normalized test feature, $S_{\text{base}}(\mathbf{x})$ is the energy score, $\mathbf{z}_{(i)} = \phi(\mathbf{x}_{(i)})/\|\phi(\mathbf{x}_{(i)})\|_2$ is the L2-normalized i-th nearest neighbor feature from the bank set, $s_{(i)} = S_{\text{base}}(\mathbf{x}_{(i)})$ is its energy score, and $\text{sim}(\cdot, \cdot)$ denotes cosine similarity. The neighbors are reordered in descending order of confidence-scaled similarity $s_{(i)} \, \text{sim}(\mathbf{z}_{(i)}, \mathbf{z})$. The bank set consists of $\alpha\%$ of features randomly sampled from the training set. Both α and k are selected via validation AUROC, with α chosen from $\{1, 3, 5\}$ and k from $\{1, 3, 5, 10\}$. The selected values are $k = 10$ / $\alpha = 1\%$ (ResNet-50), $k = 10$ / $\alpha = 5\%$ (Swin-B), $k = 10$ / $\alpha = 3\%$ (ViT-B), $k = 1$ / $\alpha = 5\%$ (CIFAR-10), and $k = 10$ / $\alpha = 5\%$ (CIFAR-100).

Computing Infrastructure. All experiments are conducted on a machine equipped with an AMD Ryzen 9 7950X CPU and an NVIDIA RTX 4090 GPU.

D Additional Experiments

D.1 Full Experimental Results by Dataset and Backbone

This section provides detailed per-dataset and per-backbone OOD detection results to complement the aggregated results presented in Section 5. Specifically, Tables 4–6 report per-dataset results on ImageNet for ResNet-50, Swin-B, and ViT-B, respectively, and Tables 7–8 report per-dataset results on CIFAR-10 and CIFAR-100. MahaVar achieves the best average performance among all Mahalanobis-based methods across all backbone architectures on ImageNet, and outperforms all evaluated baselines on average on Swin-B and ViT-B. On CIFAR-100, MahaVar achieves the best average performance across all evaluated methods, while on CIFAR-10 it remains competitive with the best-performing baselines. Notably, the performance gains on ViT-B are more modest compared to ResNet-50 and Swin-B, consistent with the insufficient penultimate layer dimension ($d = 768 < C - 1 = 999$) that prevents complete ETF formation as discussed in Section 4.

Table 4: OOD detection performance on ImageNet with ResNet-50. Results are reported in terms of AUROC and FPR@95 for each OOD dataset. The best and second best results are highlighted in **bold** and underline, respectively.

Method	Near-OOD				Far-OOD						Avg	
	SSB-Hard		NINCO		iNaturalist		Texture		OpenImage-O			
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
MSP	72.16	84.49	79.98	75.94	88.43	52.73	82.48	63.93	84.99	63.61	81.61	68.14
MaxLogit	72.75	83.70	80.41	76.61	91.15	50.82	88.42	51.30	89.26	57.13	84.40	63.91
KL-matching	68.61	82.72	79.90	73.32	89.78	44.56	84.08	64.57	86.08	60.30	81.69	65.10
Energy	72.35	83.82	79.70	77.56	90.61	53.83	88.74	49.26	89.15	57.09	84.11	64.31
GEN	72.19	85.75	81.80	76.52	92.45	45.79	87.59	57.27	89.35	60.29	84.67	65.12
ReAct	73.08	80.47	81.71	70.43	96.32	19.96	92.80	38.41	91.88	41.33	87.16	50.12
ASH-S	74.59	80.89	84.42	63.51	97.70	12.33	97.86	10.60	93.83	30.66	89.68	39.60
SCALE	**77.23**	**78.00**	**85.27**	**61.79**	**98.00**	**10.52**	97.62	11.47	**93.97**	**29.92**	**90.42**	**38.34**
KNN	60.41	91.33	77.79	77.17	83.63	63.34	96.48	14.92	83.97	60.13	80.46	61.38
NCI	66.41	88.99	80.14	77.83	92.65	46.80	93.04	37.62	90.48	54.99	84.54	61.25
NCI+filter	71.43	83.55	83.26	68.97	96.43	20.51	96.34	18.33	93.39	36.36	88.17	45.54
fDBD	70.25	86.67	82.58	74.65	93.70	40.34	93.45	34.59	91.15	52.61	86.22	57.77
ViM	65.11	90.53	78.55	77.82	89.55	59.21	97.96	9.67	90.40	51.33	84.31	57.71
NNGuide	73.67	80.61	81.09	70.30	94.50	29.16	93.87	25.58	91.57	41.36	86.94	49.40
Mahalanobis	47.34	97.52	62.46	94.81	64.16	93.81	89.71	45.08	69.01	88.10	66.54	83.86
Mahalanobis++	66.34	86.21	83.11	67.39	95.13	24.87	**98.42**	**7.79**	92.24	39.30	87.05	45.11
MahaVar (Ours)	68.72	83.59	84.44	63.90	96.90	14.57	98.37	8.02	92.77	36.25	88.24	41.27

Table 5: OOD detection performance on ImageNet with Swin-B. Results are reported in terms of AUROC and FPR@95 for each OOD dataset. The best and second best results are highlighted in **bold** and underline, respectively.

Method	Near-OOD				Far-OOD						Avg	
	SSB-Hard		NINCO		iNaturalist		Texture		OpenImage-O			
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
MSP	71.30	80.20	78.42	71.66	86.46	50.79	80.12	63.33	81.76	62.34	79.61	65.66
MaxLogit	65.22	81.38	71.87	72.58	78.89	55.39	74.68	60.27	70.92	66.17	72.32	67.16
KL-matching	71.07	80.05	80.76	69.11	90.05	45.72	84.42	59.86	87.57	55.93	82.77	62.13
Energy	59.27	86.00	64.28	79.82	67.79	73.19	69.63	64.96	60.40	76.75	64.27	76.14
GEN	74.20	78.55	82.96	64.66	92.01	34.05	86.19	50.58	87.65	48.81	84.60	55.33
ReAct	68.61	84.90	80.24	75.16	88.60	58.96	85.56	56.72	85.83	62.41	81.77	67.63
ASH-S	35.54	98.21	26.43	99.06	15.22	99.69	22.06	98.74	16.83	99.40	23.22	99.02
SCALE	55.97	92.86	62.51	89.69	63.10	91.18	69.55	79.61	66.71	84.48	63.57	87.56
KNN	71.92	85.84	83.08	72.47	92.32	48.32	90.22	45.76	91.62	49.66	85.83	60.41
NCI	70.40	84.68	81.47	72.01	91.76	47.20	87.55	55.37	90.78	49.23	84.39	61.70
NCI+filter	70.45	84.54	81.53	71.93	91.81	46.80	87.62	55.33	90.79	49.22	84.44	61.56
fDBD	69.96	85.04	81.22	72.51	91.88	47.64	88.10	55.12	91.03	48.96	84.44	61.85
ViM	74.46	82.03	84.54	69.26	94.74	33.74	88.78	52.29	92.83	41.52	87.07	55.77
NNGuide	73.13	84.10	84.93	68.22	93.58	39.85	**91.37**	**41.43**	92.31	47.32	87.06	56.19
Mahalanobis	74.62	81.88	85.08	68.73	94.55	31.21	88.38	58.80	92.76	40.91	87.08	56.31
Mahalanobis++	75.66	74.45	86.40	60.16	95.71	22.43	90.21	47.81	93.48	36.45	88.29	48.26
MahaVar (Ours)	**77.28**	**70.26**	**88.77**	**55.22**	**97.06**	**15.33**	90.22	48.08	**94.12**	**33.58**	**89.49**	**44.50**

Table 6: OOD detection performance on ImageNet with ViT-B. Results are reported in terms of AUROC and FPR@95 for each OOD dataset. The best and second best results are highlighted in **bold** and underline, respectively.

Method	Near-OOD				Far-OOD						Avg	
	SSB-Hard		NINCO		iNaturalist		Texture		OpenImage-O			
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
MSP	69.03	84.77	78.08	73.32	88.16	51.61	85.09	57.52	84.85	59.96	81.04	65.44
MaxLogit	64.45	85.72	72.39	74.21	85.25	52.38	83.79	53.91	81.61	58.82	77.50	65.01
KL-matching	68.47	83.72	81.10	72.03	90.18	49.87	86.93	60.45	87.51	59.37	82.84	65.09
Energy	59.16	88.16	66.05	78.45	79.26	64.06	81.22	55.66	76.49	65.05	72.44	70.27
GEN	69.68	86.14	81.36	72.01	92.92	39.89	89.85	48.04	89.50	51.93	84.66	59.60
ReAct	63.19	88.30	75.47	78.60	86.00	65.22	86.66	54.28	84.23	64.88	79.11	70.26
ASH-S	34.41	97.91	22.87	99.42	9.75	99.96	28.09	98.78	18.19	99.60	22.66	99.13
SCALE	56.78	88.75	61.34	79.94	73.62	69.13	78.94	56.78	72.51	67.33	68.64	72.39
KNN	66.54	88.96	82.22	77.75	90.53	58.95	90.57	50.02	89.10	61.45	83.79	67.43
NCI	66.43	87.06	80.75	74.65	90.09	54.78	87.82	54.01	88.48	58.93	82.72	65.89
NCI+filter	66.80	86.79	80.88	74.31	90.31	53.37	88.06	53.24	88.50	58.28	82.91	65.20
fDBD	65.71	87.25	80.74	74.31	90.49	53.07	88.52	52.93	89.03	57.61	82.90	65.03
ViM	69.33	79.21	84.58	61.94	95.54	25.29	90.55	**42.97**	92.03	43.28	86.41	50.54
NNGuide	68.20	87.62	83.63	73.85	92.35	49.16	**91.51**	46.12	90.45	58.20	85.23	62.99
Mahalanobis	71.42	75.69	86.44	**57.77**	95.99	18.56	89.33	50.27	92.33	**39.80**	87.10	**48.42**
Mahalanobis++	72.51	74.69	87.19	58.32	96.23	19.35	90.82	50.17	92.77	41.96	87.90	48.90
MahaVar (Ours)	**73.06**	**73.75**	**87.54**	57.98	**96.49**	**18.47**	90.84	51.40	**93.09**	40.65	**88.20**	48.45

Table 7: OOD detection performance on CIFAR-10 with ResNet-18. Results are reported in terms of AUROC and FPR@95 for each OOD dataset. The best and second best results are highlighted in **bold** and underline, respectively.

Method	Near-OOD				Far-OOD								Avg	
	TIN		CIFAR-100		MNIST		SVHN		Texture		Places365		Avg	
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
MSP	88.31	57.65	86.10	61.86	97.51	16.08	90.50	58.41	88.30	58.76	88.89	55.26	89.94	51.34
MaxLogit	88.14	46.17	84.91	53.69	99.30	3.52	90.92	45.21	87.67	48.85	89.16	43.05	90.02	40.08
KL-matching	85.52	57.47	82.83	61.80	96.82	18.10	86.53	59.37	84.63	58.87	85.67	54.59	87.00	51.70
Energy	88.22	46.16	84.96	53.83	**99.43**	**3.06**	91.04	**44.37**	87.74	48.78	89.28	42.46	90.11	**39.78**
GEN	88.65	47.39	85.79	54.51	99.18	4.90	91.18	47.31	88.42	49.61	89.51	44.62	90.45	41.39
ReAct	87.73	48.76	84.79	56.04	99.26	4.40	88.74	53.55	86.62	53.60	91.06	41.37	89.70	42.95
ASH-S	82.88	49.33	79.18	57.04	98.50	9.25	83.20	56.87	81.87	51.58	87.97	42.65	85.60	44.45
SCALE	84.12	49.04	80.34	56.89	98.87	6.28	84.89	55.86	82.71	52.93	88.73	40.74	86.61	43.62
KNN	**90.81**	49.02	**88.69**	55.11	96.34	23.52	92.09	50.99	91.81	47.68	91.13	48.07	**91.81**	45.73
NCI	89.10	51.65	87.20	56.92	94.99	33.73	89.60	57.71	91.30	48.16	90.09	50.90	90.38	49.85
NCI+filter	89.10	51.65	87.20	56.92	94.99	33.73	89.60	57.71	91.30	48.16	90.09	50.90	90.38	49.85
fDBD	88.47	52.48	86.35	57.57	90.70	47.32	86.19	61.54	89.51	50.99	90.33	50.31	88.59	53.37
ViM	89.09	54.86	87.19	56.99	88.33	97.36	**92.76**	45.64	**94.15**	**33.55**	88.91	56.50	90.07	57.48
NNGuide	88.61	**45.40**	85.27	**53.39**	99.00	5.51	89.79	48.64	88.84	46.97	89.81	42.30	90.22	40.37
Mahalanobis	87.30	63.63	85.70	64.38	86.57	98.39	89.24	64.62	92.22	44.61	87.66	63.36	88.12	66.50
Mahalanobis++	89.64	52.38	87.73	57.30	95.48	28.68	90.24	60.15	91.27	50.14	90.41	49.18	90.80	49.64
MahaVar (Ours)	90.61	45.81	87.73	54.84	98.39	8.87	85.18	83.94	92.25	44.31	**92.53**	**35.61**	91.12	45.56

Table 8: OOD detection performance on CIFAR-100 with ResNet-18. Results are reported in terms of AUROC and FPR@95 for each OOD dataset. The best and second best results are highlighted in **bold** and underline, respectively.

Method	Near-OOD				Far-OOD								Avg	
	TIN		CIFAR-10		MNIST		SVHN		Texture		Places365		Avg	
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
MSP	82.22	73.43	78.54	79.97	74.45	88.53	79.37	79.29	78.07	79.61	79.61	78.75	78.71	79.93
MaxLogit	82.74	72.88	79.20	79.40	77.82	86.21	82.33	77.22	79.36	78.72	80.06	78.70	80.25	78.85
KL-matching	82.18	72.60	77.91	79.66	75.76	83.82	80.21	75.37	78.61	75.64	79.60	77.19	79.04	77.38
Energy	82.51	72.45	79.01	79.66	78.22	84.45	82.67	76.03	79.33	77.91	79.81	79.06	80.26	78.26
GEN	83.09	72.45	**79.33**	79.50	77.11	86.25	81.87	76.91	79.51	78.46	80.36	78.62	80.21	78.70
ReAct	82.61	72.48	78.71	79.76	77.47	84.48	82.48	76.10	80.55	77.57	80.10	79.22	80.32	78.27
ASH-S	82.27	73.94	79.17	79.34	80.08	80.06	85.84	66.58	81.79	72.80	80.77	77.06	81.65	74.96
SCALE	82.53	73.51	79.24	**79.16**	79.62	81.63	85.14	68.94	81.28	74.29	80.77	76.71	81.43	75.71
KNN	81.28	76.26	75.02	82.78	76.44	84.07	78.10	78.30	81.66	70.43	77.59	80.27	78.35	78.69
NCI	83.78	72.05	78.25	81.16	78.04	87.75	82.99	76.46	84.18	71.86	81.05	77.78	81.38	77.84
NCI+filter	83.78	72.05	78.25	81.16	78.04	87.75	82.99	76.46	84.18	71.86	81.05	77.78	81.38	77.84
fDBD	**83.82**	**71.31**	78.24	80.72	77.51	87.34	81.56	77.05	82.96	72.71	80.58	77.62	80.78	77.79
ViM	77.75	82.21	70.89	87.14	**83.44**	**76.44**	80.92	80.51	87.22	56.31	74.72	84.61	79.16	77.87
NNGuide	82.94	73.83	78.60	81.90	79.70	83.12	81.54	77.06	81.79	73.05	79.78	79.55	80.72	78.08
Mahalanobis	62.90	93.96	56.45	95.72	67.81	96.46	72.81	91.93	76.96	73.16	64.18	92.85	66.85	90.68
Mahalanobis++	82.81	74.64	76.97	82.07	81.09	78.24	86.92	58.43	86.81	57.11	81.89	71.51	82.75	70.33
MahaVar (Ours)	83.48	71.56	76.42	82.25	80.88	82.20	**88.39**	**57.00**	**88.73**	**51.79**	**84.89**	**63.39**	**83.80**	**68.03**

D.2 Comparison with Recent Baselines

We measure OOD detection performance under the ImageNet ID setting against baselines not included in the main experiments: CADRef [21], LTS [7], OptFS [48], and AdaSCALE-A/L [35], reported in Table 9 through Table 11. We exclude Catalyst [11] and DAVIS [12] from this comparison due to their architecture-dependent implementation. Both methods operate on spatial pre-pooling activation maps, and although DAVIS accommodates hierarchical backbones with a GAP stage such as Swin Transformer, neither extends to architectures such as ViT, which lacks an explicit GAP operation. Table 9 shows that MahaVar trails activation-shaping methods such as AdaSCALE and logit-scaling methods such as LTS by a narrow margin in AUROC, mirroring the gap observed against SCALE in the main ResNet-50 results (Table 4). However, Table 10 and Table 11 show that MahaVar substantially outperforms these activation- and logit-scaling methods on Transformer-based backbones. A similar pattern holds against CADRef: while MahaVar trails CADRef by 1.35pp in AUROC and 1.10pp in FPR@95 on ResNet-50, it outperforms CADRef by 1.91pp and 1.61pp in AUROC on Swin-B and ViT-B respectively, and by approximately 9pp in FPR@95 on both backbones.

Table 9: OOD detection performance on ImageNet with ResNet-50 against additional recent baselines. Results are reported in terms of AUROC and FPR@95 for each OOD dataset. The best and second best results are highlighted in **bold** and underline, respectively.

Method	Near-OOD				Far-OOD						Avg	
	SSB-Hard		NINCO		iNaturalist		Texture		OpenImage-O			
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
CADRef	73.93	79.91	85.17	63.05	96.91	15.88	97.95	9.71	93.96	32.31	89.59	40.17
OptFS	71.11	83.79	82.45	70.48	96.62	18.43	96.10	21.86	92.55	39.26	87.76	46.77
AdaSCALE-A	77.96	77.85	84.13	66.52	95.26	27.95	94.30	28.84	94.04	33.49	89.14	46.93
AdaSCALE-L	**78.16**	**77.19**	85.00	63.70	96.48	19.79	95.97	20.39	**94.70**	**29.40**	90.06	42.09
LTS	77.04	78.08	**85.23**	**61.77**	**97.85**	**11.23**	97.41	12.29	94.03	30.12	**90.31**	**38.70**
MahaVar (Ours)	68.72	83.59	84.44	63.90	96.90	14.57	**98.37**	**8.02**	92.77	36.25	88.24	41.27

Table 10: OOD detection performance on ImageNet with Swin-B against additional recent baselines. Results are reported in terms of AUROC and FPR@95 for each OOD dataset. The best and second best results are highlighted in **bold** and underline, respectively.

Method	Near-OOD				Far-OOD						Avg	
	SSB-Hard		NINCO		iNaturalist		Texture		OpenImage-O			
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
CADRef	74.99	78.54	85.48	64.82	94.03	35.58	**90.84**	**44.32**	92.53	45.09	87.58	53.67
OptFS	68.90	85.30	80.56	72.90	90.82	49.66	87.15	56.55	90.08	50.97	83.50	63.08
AdaSCALE-A	69.35	82.26	79.23	70.53	86.10	54.83	79.41	58.91	82.94	62.14	79.40	65.73
AdaSCALE-L	69.29	83.18	79.05	71.96	85.14	61.89	78.43	61.59	82.33	67.01	78.85	69.13
LTS	65.41	82.51	75.66	72.54	84.82	52.99	79.66	60.91	79.49	63.65	77.01	66.52
MahaVar (Ours)	**77.28**	**70.26**	**88.77**	**55.22**	**97.06**	**15.33**	90.22	48.08	**94.12**	**33.58**	**89.49**	**44.50**

Table 11: OOD detection performance on ImageNet with ViT-B against additional recent baselines. Results are reported in terms of AUROC and FPR@95 for each OOD dataset. The best and second best results are highlighted in **bold** and underline, respectively.

Method	Near-OOD				Far-OOD						Avg	
	SSB-Hard		NINCO		iNaturalist		Texture		OpenImage-O			
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
CADRef	71.02	82.69	85.14	68.65	93.85	38.28	**91.36**	**46.40**	91.57	50.10	86.59	57.22
OptFS	66.18	87.22	80.51	74.63	89.92	56.08	88.42	53.91	88.22	60.09	82.65	66.39
AdaSCALE-A	68.03	85.81	76.48	73.80	84.77	59.27	82.71	57.73	83.45	60.77	79.09	67.48
AdaSCALE-L	66.87	87.34	75.30	76.56	79.84	74.55	78.85	64.05	80.76	67.06	76.32	73.91
LTS	64.36	85.13	75.46	72.97	88.89	46.28	86.56	49.67	83.44	59.99	79.74	62.81
MahaVar (Ours)	**73.06**	**73.75**	**87.54**	**57.98**	**96.49**	**18.47**	90.84	51.40	**93.09**	**40.65**	**88.20**	**48.45**

D.3 Effect of Variance Term and L2-Normalization

Table 12 presents an ablation study on the contribution of the variance term and L2 normalization, where the hyperparameter α for each method is selected using the OpenImage-O validation set provided by OpenOOD v1.5. Adding the variance term to the standard Mahalanobis distance without L2 normalization (Mahalanobis + Var) already yields substantial improvements over the baseline, demonstrating that class-wise distance variance carries meaningful discriminative information for OOD detection. However, Mahalanobis + Var falls short of Mahalanobis++, suggesting that L2 normalization plays an important complementary role in stabilizing the feature geometry. Notably, using the variance term in isolation (Var Only) substantially underperforms both Mahalanobis++ and MahaVar, indicating that variance alone is insufficient as a standalone OOD score. This confirms that the two terms are complementary rather than substitutable: the nearest-class distance provides a strong signal across the bulk of samples, whereas the variance term additionally recovers discriminative power that the nearest-class distance alone does not capture. MahaVar, which combines both L2 normalization and the variance term, achieves the best average performance across all datasets, confirming that the two components are mutually beneficial and that the variance term provides consistent gains on top of the strong L2-normalized Mahalanobis baseline.

Table 12: Ablation study on the effect of the variance term and L2 normalization on ImageNet benchmark. Results are averaged over three backbone architectures: ResNet-50, Swin-B, and ViT-B. The best results are highlighted in **bold**.

Method	Near-OOD				Far-OOD						Avg	
	SSB-Hard		NINCO		iNaturalist		Texture		OpenImage-O			
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
Mahalanobis	64.46	85.03	77.99	73.77	84.90	47.86	89.14	51.38	84.70	56.27	80.24	62.86
Mahalanobis + Var	72.98	78.38	84.18	66.77	96.00	22.72	90.23	48.63	90.39	45.13	86.76	52.33
Var Only (L2)	58.38	89.79	58.05	91.78	75.42	75.25	48.29	96.42	59.43	91.27	59.91	88.90
Mahalanobis++	71.50	78.45	85.56	61.95	95.69	22.22	**93.15**	**35.26**	92.83	39.24	87.75	47.42
MahaVar (Ours)	**73.02**	**75.87**	**86.92**	**59.03**	**96.82**	**16.12**	93.14	35.83	**93.33**	**36.83**	**88.64**	**44.74**

We further formalize this complementarity with a pointwise argument, showing precisely when the combined score benefits from summing the two terms. Let $d_c^2(x) \triangleq \|\hat{\phi}(x) - \mu_c\|_2^2$ denote the squared distance to class mean c, and let $d_{(1)}^2(x) \triangleq \min_c d_c^2(x)$ denote the nearest-class distance. For any $x \in \mathcal{X}$, the identity underlying the proof of Corollary 4.4 holds unconditionally, without any assumption on x:

$$\mathrm{Var}_c\big[d_c^2(x)\big] = \frac{4R^2}{C-1}\rho(x), \qquad \rho(x) \triangleq \|P\hat{\phi}(x)\|_2^2 \in [0,1], \tag{53}$$

where P denotes the orthogonal projection onto $\mathrm{span}(\mu_1, \ldots, \mu_C)$. Consider the combined score $\tilde{S}(x) \triangleq -d_{(1)}^2(x) + \alpha \cdot \mathrm{Var}_c[d_c^2(x)]$. For any pair $x_{\text{id}} \in \mathcal{X}_{\text{in}}$, $x_{\text{ood}} \in \mathcal{X} \setminus \mathcal{X}_{\text{in}}$, define

$$\Delta \triangleq d_{(1)}^2(x_{\text{id}}) - d_{(1)}^2(x_{\text{ood}}), \qquad g \triangleq \frac{4R^2}{C-1}(\rho_{\text{id}} - \rho_{\text{ood}}), \tag{54}$$

so that $\tilde{S}(x_{\text{id}}) > \tilde{S}(x_{\text{ood}}) \iff \alpha \cdot g > \Delta$. Under Theorem 4.3's hypothesis, where x_{ood} satisfies $\max_c |s_c| < R(R-\epsilon)/\sqrt{C-1}$ with $s_c \triangleq \hat{\phi}(x_{\text{ood}})^\top \mu_c$, the min-distance term already ranks the pair correctly. Writing $t \triangleq \max_c |s_c|$, the hypothesis states $t < R(R-\epsilon)/\sqrt{C-1}$. Since $s_c \le |s_c| \le t$ for every class c, we have $\max_c s_c \le t$, and therefore

$$d_{(1)}^2(x_{\text{ood}}) = 1 + R^2 - 2\max_c s_c \ge 1 + R^2 - 2t > 1 + R^2 - \frac{2R(R-\epsilon)}{\sqrt{C-1}} \ge 1 + R^2 - 2R(R-\epsilon), \tag{55}$$

where the second inequality substitutes the hypothesis bound on t and the last uses $\sqrt{C-1} \ge 1$ for $C \ge 2$. Expanding $(R-\epsilon)^2 = R^2 - 2R\epsilon + \epsilon^2$ gives $1 + R^2 - 2R(R-\epsilon) = 1 - (R-\epsilon)^2 + \epsilon^2$. Since $0 < \epsilon < R \le 1$ implies $(R-\epsilon)^2 < R^2 \le 1$, the leading term is strictly positive, so $d_{(1)}^2(x_{\text{ood}}) > \epsilon^2$. The within-class deviation bound in Assumption 4.1 directly gives $d_{(1)}^2(x_{\text{id}}) \le \epsilon^2$, and combining the two bounds yields $\Delta \le 0$. Since Theorem 4.3 also guarantees $g > 0$ in this regime, both terms agree on the correct ordering, and the variance term reinforces a comparison the min-distance term did not need to fix.

In contrast, when x_{ood} is unconstrained as in Corollary 4.4, Δ can be positive: the min-distance term can genuinely fail if x_{ood} happens to sit closer to some class mean than x_{id} does. Setting aside Corollary 4.4's assumption that $\rho_{\text{id}} = 1$ and treating $\rho_{\text{id}}, \rho_{\text{ood}}$ as general, $g > 0$ reduces to the single condition $\rho_{\text{id}} > \rho_{\text{ood}}$, independent of Δ. On average, this ordering holds in practice, with ρ_{id} exceeding ρ_{ood} across all 22 of 22 backbone and dataset combinations for which $D \geq C$, spanning every ID/OOD pair in the OpenOOD v1.5 setting. Whenever $g > 0$, a finite threshold

$$\alpha^*(x_{\text{id}}, x_{\text{ood}}) = \frac{\Delta}{g} = \frac{\Delta}{\frac{4R^2}{C-1}(\rho_{\text{id}} - \rho_{\text{ood}})} \tag{56}$$

exists such that $\alpha > \alpha^*$ restores the correct ordering $\tilde{S}(x_{\text{id}}) > \tilde{S}(x_{\text{ood}})$, even when $\Delta > 0$. This is the precise sense in which the two terms are complementary by construction: they fail under different, non-overlapping geometric conditions, and $\alpha \cdot g > \Delta$ is exactly the condition under which the variance term compensates for a min-distance failure.

This ordering can be understood through the identity itself: since $\rho(x) \in [0, 1]$, class-wise distance variance grows monotonically in ρ from 0, corresponding to a sample orthogonal to every class mean, up to the ETF-geometry maximum at $\rho = 1$. Corollary 4.4's assumption $\rho_{\text{id}} = 1$ sets the top of this range, so it should be read as the idealized limit of the identity rather than a claim we expect every ID sample to satisfy exactly. What the assumption isolates is the mechanism rather than the exact separation: since the same identity and the same constants R, C apply to both ID and OOD samples, the variance separation the theory predicts reduces entirely to the ordering $\rho_{\text{id}} > \rho_{\text{ood}}$, which is precisely the empirical regularity verified above.

This mechanism also explains why the Var Only ablation in Table 12 underperforms both Mahalanobis++ and MahaVar. Since $\rho(x)$ captures only the shape of a sample's alignment with the class-mean subspace and not the magnitude of its distance to the nearest class, using ρ alone can misread a sample that is unambiguously far from every class in raw distance if that sample happens to align well with the subspace. This is a genuine limitation of variance as a standalone score, and it is exactly why the combined score falls back on the min-distance term rather than relying on variance alone.

D.4 Effect of Higher-Order Moment Terms

To further investigate the discriminative structure of class-wise Mahalanobis distances, we explore incorporating the skewness of the distance distribution as an additional signal. Specifically, we extend the MahaVar score as follows:

$$S(x) = -\min_c d_{\text{Maha},c}(\hat{\phi}(x)) + \alpha \cdot \text{Var}_c[d_{\text{Maha},c}(\hat{\phi}(x))] + \beta \cdot \text{Skew}_c[d_{\text{Maha},c}(\hat{\phi}(x))] \tag{57}$$

where $\beta \in \mathbb{R}$ controls the contribution and sign of the skewness term, selected via validation AUROC. As shown in Table 13, incorporating skewness yields marginal but consistent improvements over MahaVar across most datasets, suggesting that the asymmetric structure of the distance distribution contains additional discriminative information beyond variance alone. Notably, the optimal sign of β varies across backbone architectures: a positive β is selected for ResNet-50, while a negative β is selected for Swin-B and ViT-B, reflecting differences in the skewness structure of class-wise distances across feature geometries. However, since the gains are modest and the optimal sign of β is architecture-dependent, we adopt the variance-only formulation as our main method.

Table 13: Ablation study on the effect of the skewness term on ImageNet benchmark. Results are averaged over three backbone architectures: ResNet-50, Swin-B, and ViT-B. The selected β values are 10 (ResNet-50), -10 (Swin-B), and -10 (ViT-B). The best results are highlighted in **bold**.

Method	SSB-Hard		NINCO		iNaturalist		Texture		OpenImage-O		Avg	
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
Mahalanobis++	71.50	78.45	85.56	61.95	95.69	22.22	93.15	35.26	92.83	39.24	87.75	47.42
MahaVar (Ours)	73.02	75.87	86.92	59.03	96.82	16.12	93.14	35.83	93.33	36.83	88.64	44.74
Mahalanobis++ + Skew	71.58	78.46	85.70	61.89	95.71	22.10	**93.20**	**35.06**	92.85	39.12	87.81	47.33
MahaVar + Skew (Ours)	**73.13**	**75.85**	**87.07**	**58.84**	**96.85**	**15.89**	**93.20**	35.54	**93.36**	**36.70**	**88.72**	**44.56**

D.5 Hyperparameter Sensitivity Analysis

We analyze the sensitivity of MahaVar to the hyperparameter α controlling the contribution of the variance term. The analysis is conducted on the ImageNet validation set provided by OpenOOD v1.5,

using in-distribution ImageNet validation samples and OOD OpenImage-O validation samples. The results are reported in Table 14. As α increases from zero, performance improves steadily across all backbone architectures, but eventually degrades as the variance term begins to dominate the nearest-class distance term. The optimal α lies in the range of 0.05 to 0.1 across all backbones, and we recommend practitioners to search within this range for reliable performance.

Table 14: Sensitivity analysis of the hyperparameter α on the ImageNet validation set (OpenImage-O), reported in terms of AUROC. The best performance for each backbone is highlighted in **bold**.

	α																		
Backbone	0	0.0001	0.0003	0.0005	0.001	0.002	0.003	0.005	0.007	0.01	0.012	0.015	0.02	0.03	0.05	0.07	0.1	0.15	0.2
ResNet-50	92.25	92.26	92.26	92.26	92.28	92.30	92.32	92.36	92.40	92.45	92.49	92.53	92.60	92.69	**92.75**	92.64	92.20	90.95	89.30
Swin-B	93.35	93.35	93.36	93.36	93.37	93.38	93.39	93.42	93.45	93.49	93.51	93.55	93.61	93.72	93.90	94.02	**94.10**	93.97	93.54
ViT-B	92.28	92.28	92.29	92.29	92.29	92.30	92.31	92.32	92.33	92.35	92.37	92.38	92.41	92.46	92.52	**92.55**	92.50	92.24	91.73

D.6 Effect of Distance Metric

Table 15 reports the performance of MahaVar under different distance metrics. Across all metrics, incorporating the variance term consistently improves performance, confirming that class-wise distance variance provides meaningful discriminative information regardless of the choice of distance function. Among all combinations, MahaVar with the Mahalanobis distance achieves the best overall performance. We attribute this to the fact that the Mahalanobis distance accounts for the covariance structure of the feature space, providing a more geometrically faithful measure of class proximity than simple L1 or L2 distances, which in turn leads to the most discriminative OOD detection performance.

Table 15: Ablation study on the effect of distance metric on ImageNet benchmark. Results are averaged over three backbone architectures: ResNet-50, Swin-B, and ViT-B. The best results are highlighted in **bold**.

Method	Near-OOD				Far-OOD						Avg	
	SSB-Hard		NINCO		iNaturalist		Texture		OpenImage-O			
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
L1	70.61	81.10	84.26	66.67	94.23	31.27	92.53	38.20	91.70	44.88	86.67	52.42
L1 + Var	72.12	77.56	85.36	63.49	95.18	26.03	92.76	38.80	92.10	43.48	87.50	49.87
L2	71.40	79.83	84.79	64.69	95.35	24.88	92.99	35.38	92.42	41.58	87.39	49.27
L2 + Var	72.37	77.68	85.55	62.18	96.06	20.62	**93.31**	**34.41**	92.83	39.41	88.02	46.86
Mahalanobis++	71.50	78.45	85.56	61.95	95.69	22.22	93.15	35.26	92.83	39.24	87.75	47.42
MahaVar (Ours)	**73.02**	**75.87**	**86.92**	**59.03**	**96.82**	**16.12**	93.14	35.83	**93.33**	**36.83**	**88.64**	**44.74**

D.7 Effect of Top-K Nearest Classes for Variance Computation

Table 16 reports the AUROC of MahaVar as a function of the number of nearest classes K used for variance computation, evaluated on the ImageNet benchmark across three backbone architectures. While the sharp minimum structure observed in Section 3 suggests that the distance structure near the closest class already carries rich discriminative information, our experiments reveal that incorporating distances to a larger number of classes generally improves performance. When K is small, the variance term provides little additional discriminative information beyond the nearest-class distance, as reflected by $\alpha = 0$ being selected in many settings. This effect is particularly pronounced for ResNet-50, where $\alpha = 0$ is selected for all $K \leq 100$, whereas using all classes ($K = 1000$) leads to substantial AUROC gains. Across all three backbones, AUROC generally improves as K increases, suggesting that incorporating the distance structure of a larger number of classes enriches the variance signal. The exception is ViT-B, where $K = 500$ achieves the best average AUROC rather than $K = 1000$, which we attribute to the insufficient penultimate layer dimension ($d = 768 < C - 1 = 999$) that prevents complete ETF formation as discussed in Section 4, causing the distances to farther classes to lose their geometric regularity and degrade the quality of the variance signal. Nevertheless, we report results with $K = 1000$ across all backbones including ViT-B in the main experiments, following a consistent evaluation protocol.

Table 16: Ablation study on the number of nearest classes K used for variance computation on ImageNet, reported in terms of AUROC, evaluated across ResNet-50, Swin-B, and ViT-B. The best α for each setting is selected on the validation set, and the best results for each backbone are highlighted in **bold**.

Backbone	K	SSB-Hard	NINCO	iNaturalist	Texture	OpenImage-O	Avg	Best α
ResNet-50	5	66.34	83.11	95.13	**98.42**	92.24	87.05	0.0000
	10	66.34	83.11	95.13	**98.42**	92.24	87.05	0.0000
	20	66.34	83.11	95.13	**98.42**	92.24	87.05	0.0000
	30	66.34	83.11	95.13	**98.42**	92.24	87.05	0.0000
	50	66.34	83.11	95.13	**98.42**	92.24	87.05	0.0000
	100	66.34	83.11	95.13	**98.42**	92.24	87.05	0.0000
	200	66.25	83.02	95.16	98.39	92.25	87.02	0.0030
	400	66.33	82.79	95.61	98.28	92.36	87.07	0.0300
	500	66.69	83.05	95.73	98.31	92.40	87.24	0.0300
	all (1000)	**68.72**	**84.44**	**96.90**	98.37	**92.77**	**88.24**	0.0500
Swin-B	5	75.66	86.40	95.71	90.21	93.48	88.29	0.0000
	10	75.25	86.21	95.68	90.25	93.49	88.17	0.0010
	20	72.87	85.26	95.76	90.64	93.68	87.64	0.0200
	30	72.05	84.99	95.96	90.90	93.85	87.55	0.0500
	50	71.67	84.89	96.29	91.13	94.05	87.61	0.1000
	100	71.53	84.90	96.79	**91.26**	94.28	87.75	0.2000
	200	72.58	85.77	97.15	91.01	94.27	88.16	0.2000
	400	74.81	87.30	97.30	90.66	94.22	88.86	0.1500
	500	75.56	87.96	**97.50**	90.58	**94.31**	89.18	0.1500
	all (1000)	**77.28**	**88.77**	97.06	90.22	94.12	**89.49**	0.1000
ViT-B	5	72.51	87.19	96.23	90.82	92.77	87.90	0.0000
	10	72.37	87.12	96.21	90.83	92.77	87.86	0.0005
	20	71.12	86.53	96.14	90.99	92.82	87.52	0.0120
	30	70.56	86.28	96.20	91.06	92.87	87.40	0.0300
	50	70.56	86.33	96.39	**91.08**	92.96	87.46	0.0500
	100	70.66	86.46	96.73	91.01	93.07	87.59	0.1000
	200	71.23	86.87	97.04	90.78	93.09	87.80	0.1500
	400	72.39	87.62	97.02	90.74	93.10	88.17	0.1000
	500	72.68	**88.00**	**97.34**	90.62	**93.21**	**88.37**	0.1500
	all (1000)	**73.06**	87.54	96.49	90.84	93.09	88.20	0.0700

D.8 Effect of Feature Centering in L2-Normalized Mahalanobis Detection

In Theorem 4.3 and Corollary 4.4, the theoretical analysis is conducted with $\hat{\phi}(\mathbf{x}) = (\phi(\mathbf{x}) - \boldsymbol{\mu}_G)/\|\phi(\mathbf{x}) - \boldsymbol{\mu}_G\|_2$, i.e., features are centered by the global mean $\boldsymbol{\mu}_G$ prior to L2 normalization, so that the simplex ETF structure applies directly to the class means $\boldsymbol{\mu}_c$ in the normalized space. To assess whether this centering step is beneficial in practice, we compare centered and non-centered variants of both Mahalanobis and MahaVar across all three backbone architectures on ImageNet, with results reported in Table 17. For Swin-B and ViT-B, centering yields nearly identical performance to the non-centered variant ($\|\boldsymbol{\mu}_G\|_2 = 2.03$ and 1.67, respectively), suggesting that the practical impact of centering is limited in these architectures. For ResNet-50, however, centering leads to a substantial performance degradation, which we attribute to the non-negative feature geometry induced by the terminal ReLU activation, reflected in its substantially larger global mean norm ($\|\boldsymbol{\mu}_G\|_2 = 21.5$). Despite the theoretical motivation for centering, we adopt the non-centered L2 normalization in the practical implementation of MahaVar, and leave a full theoretical treatment of the $\boldsymbol{\mu}_G \neq \mathbf{0}$ setting as well as a deeper understanding of the performance degradation observed for ResNet as future work.

Table 17: Ablation study on the effect of feature centering on ImageNet benchmark. Mahalanobis++ (ctr) and MahaVar (ctr) denote variants where the global mean $\boldsymbol{\mu}_G$ is subtracted prior to L2 normalization, i.e., $\hat{\phi}(\mathbf{x}) = (\phi(\mathbf{x}) - \boldsymbol{\mu}_G)/\|\phi(\mathbf{x}) - \boldsymbol{\mu}_G\|_2$. Results are reported in terms of AUROC and FPR@95 across three backbone architectures. The best results for each backbone are highlighted in bold.

Method	SSB-Hard		NINCO		iNaturalist		Texture		OpenImage-O		Avg	
	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95
ResNet-50												
Mahalanobis++	66.34	86.21	83.11	67.39	95.13	24.87	**98.42**	**7.79**	92.24	39.30	87.05	45.11
MahaVar	68.72	83.59	**84.44**	**63.90**	**96.90**	**14.57**	98.37	8.02	**92.77**	**36.25**	**88.24**	**41.27**
Mahalanobis++ (ctr)	63.50	90.86	77.76	80.08	79.79	86.29	89.26	48.55	83.20	70.65	78.70	75.28
MahaVar (ctr)	**70.02**	**79.28**	81.13	70.23	94.99	29.82	92.29	40.12	88.14	56.44	85.31	55.18
Swin-B												
Mahalanobis++	75.66	74.45	86.40	60.16	95.71	22.43	90.21	47.81	93.48	36.45	88.29	48.26
MahaVar	**77.28**	70.26	**88.77**	55.22	**97.06**	15.33	90.22	48.08	**94.12**	33.58	89.49	44.50
Mahalanobis++ (ctr)	75.94	73.33	86.69	59.22	95.98	20.95	90.01	49.38	93.56	36.17	88.44	47.81
MahaVar (ctr)	77.19	**70.11**	88.73	**54.92**	97.05	**15.18**	**90.46**	**46.82**	**94.12**	**33.21**	**89.51**	**44.05**
ViT-B												
Mahalanobis++	72.51	74.69	87.19	58.32	96.23	19.35	90.82	**50.17**	92.77	41.96	87.90	48.90
MahaVar	**73.06**	73.75	**87.54**	**57.98**	**96.49**	**18.47**	**90.84**	51.40	**93.09**	**40.65**	**88.20**	**48.45**
Mahalanobis++ (ctr)	72.64	74.42	87.30	58.03	96.24	19.26	90.83	50.21	92.76	42.03	87.95	48.79
MahaVar (ctr)	73.00	**73.66**	87.48	58.12	96.44	18.64	90.77	51.84	93.05	40.89	88.15	48.63

D.9 A Hyperparameter-Free Scoring

The additive score in Equation 9 combines the nearest-class distance and the variance term through a hyperparameter α, which requires an OOD validation set for tuning. We therefore also consider a ratio-based variant that removes this dependency entirely:

$$S_{\text{ratio}}(x) \triangleq \frac{\text{Var}_c[d_{\text{Maha},c}(x)]}{d_{(1)}(x)}. \tag{58}$$

This method is motivated by the relationship that ID samples tend to exhibit a smaller nearest-class distance $d_{(1)}(x)$ while OOD samples tend to exhibit the opposite, whereas the variance term exhibits the reverse tendency. We evaluate this score in Table 18, comparing it against Mahalanobis++ and MahaVar on ImageNet with ResNet-50.

Table 18: AUROC (%) of the hyperparameter-free Var/$d_{(1)}$ variant against Mahalanobis++ and MahaVar on ImageNet with ResNet-50. The best results are highlighted in **bold**.

Method	Near-OOD		Far-OOD			Avg AUROC	Avg FPR@95
	SSB-Hard	NINCO	iNaturalist	Texture	OpenImage-O		
Mahalanobis++	66.34	83.11	95.13	**98.42**	92.24	87.05	45.11
MahaVar	68.72	**84.44**	96.90	98.37	**92.77**	88.24	41.27
Var/$d_{(1)}$ (α-free)	**70.60**	84.33	**98.14**	97.14	91.81	**88.40**	**39.96**

Var/$d_{(1)}$ performs on par with the tuned additive form and in several cases marginally better, despite requiring no hyperparameter search. However, we suspect that a ratio form is inherently harder to analyze rigorously than an additive form: the additive score decomposes cleanly into the two terms Theorem 4.3 and Corollary 4.4 characterize separately, whereas the ratio couples them nonlinearly, and establishing an analogous ID/OOD separation guarantee for it would require a substantially different derivation than the one we present here. The ratio is also potentially prone to numerical instability as $d_{(1)}(x)$ approaches zero, a regime our current framework does not characterize. We view Var/$d_{(1)}$ as a promising direction and leave a full theoretical treatment to future work.

D.10 Relation to Prior Class-wise Distance Structure Observations

The general intuition that ID samples cluster closer to their own class centroid while OOD samples remain comparably distant from all centroids has been discussed in prior work, including NCDD [34]. However, prior work characterizes this asymmetry primarily through the OOD side, namely that OOD samples maintain roughly similar distances across all class centroids, without establishing the corresponding structure on the ID side. Specifically, whether the non-nearest distances plateau immediately starting from the second-nearest class, as shown in Figure 1, is a distinct claim from the magnitude of the nearest distance alone and has not been previously established.

To verify that this rank-wise plateau structure carries a genuinely different signal beyond the nearest-class distance, we test two controlled alternatives under the same L2-normalized Mahalanobis features and validation protocol as MahaVar. The first adapts NCDD's core scoring idea into our pipeline by combining the nearest distance with the sum of all non-nearest distances:

$$S(x) = -d_{(1)}(x) + \beta \cdot \sum_{c \neq \arg\min_c d_c(x)} d_c(x). \tag{59}$$

The second restricts this combination to only the nearest and second-nearest distances:

$$S(x) = -d_{(1)}(x) + \beta \left(d_{(2)}(x) - d_{(1)}(x) \right). \tag{60}$$

For both variants, β is selected via the same grid search protocol used for α in MahaVar, on ImageNet with ResNet-50. In both cases, the validation-selected β converges to 0, and the resulting scores are numerically identical to Mahalanobis++ (AUROC 87.05 / FPR@95 45.11). This indicates that neither the sum of all non-nearest distances nor the gap between the two closest classes alone carries information beyond the nearest-class distance already exploited by Mahalanobis++. In contrast, the variance term, which aggregates information across all classes rather than one or two, is selected with a non-trivial weight ($\alpha = 0.05$) and yields a clear improvement (AUROC 88.24 / FPR@95 41.27). This confirms that the discriminative signal identified in Section 3.2 depends on the full class-wise distance structure rather than any single pairwise comparison among the nearest classes.

E Limitation

MahaVar does not achieve the best overall performance on CIFAR-10, where KNN attains a higher AUROC, suggesting that non-parametric approaches may be more effective in settings with few classes. Furthermore, the gains of MahaVar are consistent on average but not uniform across individual OOD datasets, as shown by Texture on ResNet-50 and SVHN on CIFAR-10. In addition, MahaVar combines two complementary signals, and the variance term alone is not sufficient as a standalone score. The hyperparameter α is also selected on a held-out OOD validation split, which is a strong assumption that rarely holds in deployment, and alternatives such as Gaussian noise or self-generated proxy data remain unexplored. Lastly, the theoretical analysis in this work is grounded in Neural Collapse geometry, which assumes well-trained networks operating near the terminal phase of training. Relaxing these assumptions to accommodate more general feature geometries would yield a broader theoretical framework applicable to a wider range of practical settings, and we leave this as a direction for future work.

F Broader Impacts

This work proposes a post-hoc OOD detection method aimed at improving the reliability of deep neural networks in safety-critical applications. By enabling models to better identify OOD inputs, the

proposed method can contribute to safer deployment of AI systems in domains such as autonomous driving, medical imaging, and industrial inspection. We foresee no significant negative societal impacts from this work.

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 explicitly state the three main contributions of the paper, and the scope is consistently aligned with the OOD detection task throughout.

   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 limitations of this work are discussed in Appendix E.

   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 full proofs of all theoretical results are provided in Appendix B.

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 experimental settings are described in Section 5, and method details including hyperparameter configurations are provided in Appendix C.

   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 code is provided in the supplementary materials.

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: Experimental settings and details are described in Section 5 and Appendix C.

   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: [No]

   Justification: MahaVar is a deterministic post-hoc method that requires no training and involves no stochastic components; given fixed pretrained features and class statistics, all results are fully reproducible without variance across runs.

   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: The computing environment used in the experiments, including the CPU (AMD Ryzen 9 7950X) and GPU (RTX 4090), is specified in Appendix C.

   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 research conducted in this paper conforms with the NeurIPS Code of Ethics.

   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: A broader impact statement is provided in Appendix F.

    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 poses no such risks, as it does not release pre-trained models, scraped datasets, or any components with high risk for misuse.

    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: All datasets and code libraries used in this paper are properly cited.

    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 code is provided as an anonymized zip file in the supplementary materials.

    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: This paper does not involve crowdsourcing or research with human subjects.

    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: This paper does not involve crowdsourcing or research with human subjects.

    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: Large language models were used solely for writing assistance and editing purposes and did not impact the core methodology, scientific rigor, or originality of the research.

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.