Strengthen Out-of-Distribution Detection via Adaptive Mahalanobis Gap

Kunpeng Sui
School of Information Science
and Engineering, University of Jinan
skp@stu.ujn.edu.cn

Rundong He*
Department of Computing
The Hong Kong Polytechnic University
herundong0@gmail.com

Jie Su*
School of Information Science
and Engineering, University of Jinan
ise_suj@ujn.edu.cn

Abstract

Deep neural networks have achieved significant success on in-distribution (ID) tasks, but out-of-distribution (OOD) samples during inference can lead to unreliable predictions. To address this issue, distance-based methods exploit the geometric structure of feature space for OOD detection. However, methods based on static information geometry fail to address geometry distorted by ill-distributed samples. Although prior work reduces the misclassification of OOD samples as ID by adjusting the residual space with real-time features, it fails to address the misclassification of ID samples as OOD. In addition, existing distance-based methods ignore class-specific characteristics when computing deviation feature and rely on the nearest prototype for uncertainty estimation, leading to suboptimal OOD detection performance. To address these issues, we propose Adaptive Mahalanobis Gap (AMG). Specifically, we first strengthen the principal space along real-time feature directions to reduce the misclassification of ID samples as OOD. Subsequently, we incorporate class variance and prototype norms to mitigate bias of deviation feature from class-wise distributional differences. Finally, we introduce the Mahalanobis gap scoring for uncertainty scoring to overcome bias from single-prototype reliance. Experiments on CIFAR-10/100 and ImageNet-1k show that AMG consistently outperforms state-of-the-art methods in FPR@95 and AUROC. Our code is available at https://github.com/S-Skp/AMG.

1 Introduction

Deep neural networks have achieved remarkable success on in-distribution (ID) tasks, but their performance typically relies on the data distribution observed during training. When test samples deviate from the assumed distribution, the model's predictions may become unreliable, and high confidence scores may fail to ensure accurate discrimination Zhang et al. [48], Nguyen et al. [34]. Out-of-distribution (OOD) detection aims to determine whether a test sample originates from the known data distribution assumed during training [15]. OOD samples often differ from the training distribution at the semantic level, while remaining statistically similar to ID data [9]. As a result, reliable OOD detection is crucial for safety-critical applications such as autonomous driving [8], medical image analysis [10], and industrial inspection [26].

To address these challenges, researchers have recently proposed various paradigms for OOD detection. Among these, test-time OOD detection methods are well suited to pre-trained models and open-world scenarios, as they operate solely at inference time and require neither retraining nor additional supervision. Existing methods extract OOD information from multiple aspects, including model outputs [15, 13], gradient differences [18, 7], and the structure of feature representations [38, 45, 21, 27, 40, 23, 22]. Consequently, distance-based methods based on feature representations have been widely adopted.

*Corresponding authors

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

Prior work has shown that the Mahalanobis distance [37, 21] exhibits remarkable discriminative capability in OOD detection tasks. These methods measure deviations between sample features and class means while explicitly incorporating the information geometry of the training distribution, thus enhancing discrimination between ID and OOD samples. Several studies have proposed various modifications to the Mahalanobis distance to improve its robustness under complex distribution shifts, including RMD [36], Mahalanobis++ [32], and DCC [11]. RMD leverages the background distribution of the entire training data to address the failure of the Mahalanobis distance in detecting near OOD samples. Mahalanobis++ improves the robustness of Mahalanobis distance-based OOD detection by applying ℓ_2 normalization to features, mitigating the effect of feature norm variations that violate the Gaussian assumption. DCC alleviates the failure of static Mahalanobis distance under distorted information geometry by dynamically calibrating the residual-space prior covariance based on real-time features.

Although the above methods demonstrate strong empirical performance, they still fail to fully exploit the geometric structure of the feature subspace. First, existing methods [11] mitigate the misclassification of OOD samples as ID in traditional static Mahalanobis distance methods by adjusting the residual subspace with real-time input features to address for distorted information geometry. However, these methods fail to consider the possibility of misclassifying ID samples as OOD. Moreover, existing Mahalanobis distance-based methods [21, 36, 32, 11] overlook inter-class differences, limiting the discriminative capability of deviation features. Furthermore, these methods estimate uncertainty solely based on the nearest class prototype, which focuses only on local information, failing to capture the global inter-class structure.

To address the aforementioned issues, we propose the Adaptive Mahalanobis Gap (AMG). First, we adaptively reinforce the principal subspace along real-time feature directions, dynamically correcting geometric shifts during inference to maintain the consistency of ID samples. Second, we adaptively adjust the deviation features by incorporating class-wise variance and prototype norm information. This mitigates biases caused by ill-distributed classes and enables the metric to capture distributional differences across classes. Finally, we introduce the Mahalanobis Gap Scoring (MGS). It mitigates bias in uncertainty estimation caused by reliance on a single prototype. By quantifying distance differences between a test sample and all class prototypes, MGS enables more robust and discriminative uncertainty scores.

Our contributions can be summarized as below:

- We dynamically reinforce the principal subspace along real-time feature directions to correct test-time geometric shifts, reducing ID samples misclassified as OOD.
- We adjust deviation features by incorporating class-wise variance and prototype norm information, mitigating biases from ill-distributed classes and enabling the metric to capture distributional differences across classes.
- We introduce Mahalanobis Gap Scoring, which quantifies distance differences between a sample and all prototypes to provide global geometric comparison, mitigating uncertainty bias caused by over-reliance on a single prototype.
- We evaluate our method on the CIFAR and ImageNet-1k benchmarks, and the results show that our method outperforms state-of-the-art (SOTA) methods in most cases.

2 Test-time OOD Detection

Test-time OOD detection aims to identify out-of-distribution (OOD) samples during inference while keeping model parameters fixed and without requiring retraining [24]. This setting is highly compatible with post-hoc OOD detection [30], enabling training-free deployment and making it particularly suitable for large-scale pretrained models. As a result, it has attracted considerable attention in recent years.

Existing test-time OOD detection methods can be broadly categorized into six groups: logit-based [13, 15], gradient-based [3, 7], feature-based [45, 38], density-based [35, 31], subspace-based [2, 43], and distance-based [21] methods. Distance-based approaches, which detect OOD samples by measuring distances between features and in-distribution (ID) prototypes, are widely used. Among these, Mahalanobis distance–based methods, such as Mahalanobis [21], RMD [36], Mahalanobis++ [32], and DCC [11], have become particularly prominent in the field.

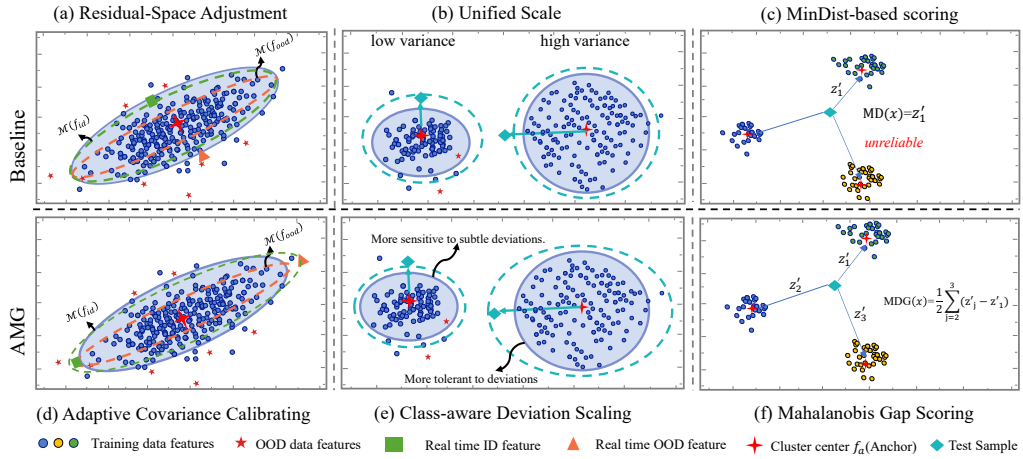


Figure 1: Visualization of the differences between the baseline and the proposed AMG framework. The top row (a-c) illustrates the limitations of the baseline in residual-space adjustment, unified scaling, and MinDist-based scoring. The bottom row (d-f) demonstrates the improvements introduced by AMG, including (d) Adaptive Covariance Calibrating, (e) Class-aware Deviation Scaling, and (f) Mahalanobis Gap Scoring.

Methods based on static information geometry fail to fully address distortions caused by poorly distributed samples, and residual-subspace adjustments alone may still misclassify ID samples. Moreover, many approaches overlook class-specific characteristics and depend only on the nearest prototype. These limitations highlight the need for methods that leverage the discriminative structure of frozen features to generate reliable test-time scores.

3 Preliminary and Insight

3.1 Problem Formulation

For out-of-distribution (OOD) detection, we consider a d-dimensional input space $\mathcal{X} \subset \mathbb{R}^d$ and label space $\mathcal{Y} = \{1, \dots, C\}$. In-distribution (ID) and OOD samples follow joint distributions $\mathcal{P}_{in}$ over $\mathcal{X}_{in} \times \mathcal{Y}_{in}$ and $\mathcal{P}_{out}$ over $\mathcal{X}_{out} \times \mathcal{Y}_{out}$, with $\mathcal{Y}_{in} \cap \mathcal{Y}_{out} = \emptyset$.

Let $F = g \circ h$ be a pre-trained network on the ID dataset $\mathcal{D}_{in} = \{(x_i, y_i)\}_{i=1}^{N}$, kept frozen during inference, where $h : \mathcal{X} \to \mathbb{R}^D$ is the feature encoder and g the classifier. For a test sample x, the feature is $f = h(x)$, producing ID and OOD feature distributions $\mathcal{F}_{in} = h(\mathcal{X}_{in})$ and $\mathcal{F}_{out} = h(\mathcal{X}_{out})$.

Distance-based OOD detection relies on the principle that for an anchor $f_a \in \mathcal{F}_{in}$,

$$d(f_{in}, f_a) \leq d(f_{out}, f_a), \tag{1}$$

where $d(\cdot, \cdot)$ is a distance metric, $f_{in} \in \mathcal{F}_{in}$, and $f_{out} \in \mathcal{F}_{out}$.

3.2 Preliminaries

Vanilla Mahalanobis Distance. The Mahalanobis distance (MD) was first applied to OOD detection in the method [21], assuming that deep features follow class-conditional Gaussian distributions. Unlike Euclidean distance, MD incorporates the covariance matrix Σ, capturing correlations and scale across feature dimensions, and providing a more reliable measure of sample deviation under non-isotropic distributions. For a test sample x and class c, it is defined as

$$d_M(f, f_a) = \sqrt{(f - f_a)^\top \Sigma^{-1} (f - f_a)}, \tag{2}$$

where f is the sample feature, f_a is the class prototype, and Σ the tied covariance matrix.

Dynamic Mahalanobis Distance. Previous OOD detection methods [21, 36, 32, 43, 2] rely on static geometry and are vulnerable to geometry distortions from ill-distributed samples.

To address this issue, DCC dynamically calibrates the covariance matrix using real-time features, enhancing sensitivity along the feature direction while constraining adjustments to the residual subspace to preserve ID structure:

$$\mathcal{M}(f) = (\Sigma - \mathbf{B}^\top \mathbf{a}\mathbf{a}^\top \mathbf{B})^{-1}, \quad f_r = \sum_i a_i b_i = \mathbf{a}^\top \mathbf{B}, \tag{3}$$

where $\mathbf{B} = \{b_i\}$ is the basis matrix of the residual subspace, $\mathbf{a} = \{a_i\}$ is the coefficient vector, and f_r represents the projection of the current feature f onto the residual subspace. Given the real-time feature f, the MD can be defined as:

$$d_{\mathcal{M}}(f, f_a) = \sqrt{(f - f_a)^\top \mathcal{M}(f)(f - f_a)}, \tag{4}$$

where f_a is the class prototype and $\mathcal{M}(f)$ is defined in Eq. (3).

3.3 Problem Analysis

Although Mahalanobis-based methods have shown success, they face structural limitations in OOD detection: (i) relying on static covariance geometry that cannot adapt to test-time features, (ii) computing deviation features in a class-agnostic manner, and (iii) depending solely on the nearest class prototype, which ignores overall inter-class distances and can misclassify samples deviating from known distributions.

To address these limitations, we propose **Adaptive Mahalanobis Gap (AMG)**, a distance-based OOD detection framework for frozen representations. As illustrated in Figure 1, AMG enhances discrimination in the principal subspace, incorporates inter-class distributional variations to alleviate uniform-scaling bias, and integrates Mahalanobis distance information to improve scoring reliability beyond the minimum-distance criterion.

4 Methodology

AMG employs a distance-based framework that incorporates: (i) Adaptive Covariance Calibrating (ACC) module to calibrate the covariance matrix with real-time features, (ii) Class-aware Deviation Scaling (CDS) module to account for intra-class variance and prototype norms, and (iii) Mahalanobis Gap Scoring (MGS) module to capture relative relationships between test data and ID classes.

4.1 Adaptive Covariance Calibrating

The effectiveness of Mahalanobis distance (MD) in distance-based OOD detection critically depends on whether the covariance matrix accurately captures the geometric structure of in-distribution (ID) data. Traditional Mahalanobis methods [21, 36, 32] rely on covariances estimated during training, which remain fixed at inference. Dynamic approaches, such as DCC [11], address this issue by adjusting the covariance in the residual subspace based on real-time features. However, it does not reinforce the principal discriminative directions, which may lead to underestimated deviations along these directions and reduced discriminative ability. To address this, we propose ACC, which dynamically adjusts the covariance matrix for each test sample while preserving the core discriminative structure of ID data.

Specifically, we eigendecompose the training covariance and form two row-stacked basis matrices: $A \in \mathbb{R}^{K \times D}$ for the top-K principal eigenvectors and $B \in \mathbb{R}^{(D-K) \times D}$ for the remaining residual eigenvectors. The principal subspace captures discriminative ID structure, whereas the residual subspace mainly contains noisy or non-informative variations.

For a real-time feature $f \in \mathbb{R}^D$, we first compute its coefficient vectors in the two subspaces:

$$p = Af, \qquad q = Bf, \tag{5}$$

where $p \in \mathbb{R}^K$ and $q \in \mathbb{R}^{D-K}$ denote the principal and residual coefficients, respectively. We then map these coefficients back to the original feature space:

$$f_A = A^\top p, \qquad f_B = B^\top q, \tag{6}$$

where $f_A, f_B \in \mathbb{R}^D$ denote the principal and residual components of f, respectively. The calibrated inverse covariance is constructed as

$$\hat{M}(f) = \left(\Sigma + \alpha f_A f_A^\top - \beta f_B f_B^\top\right)^{-1}, \tag{7}$$

where $\alpha > 0$ and $\beta > 0$ control principal-subspace expansion and residual-subspace contraction.

Proposition 1. *Let Σ be the training covariance, and $\mathbf{A}, \mathbf{B}$ its principal and residual subspaces. For any vector v and test feature f, define $\Sigma'(f) = \Sigma + \alpha f_A f_A^\top - \beta f_B f_B^\top$. Then, $v^\top \Sigma v - \beta (v^\top f_B)^2 \leq v^\top \Sigma'(f) v \leq v^\top \Sigma v + \alpha (v^\top f_A)^2$.*

Proposition 1 shows that ACC provides bounded, direction-aware covariance calibration in both principal and residual subspaces. It limits excessive adjustment along discriminative directions while controlling residual contraction, thereby reducing both OOD samples being assigned high ID confidence and ID samples being falsely rejected.

The calibrated Mahalanobis distance is computed as

$$d_{\hat{\mathcal{M}}}(f, f_a) = \sqrt{(f - f_a)^\top \hat{\mathcal{M}}(f)(f - f_a)}, \tag{8}$$

where f_a denotes the corresponding class prototype. This sample-dependent distance is later combined with class-aware deviation scaling.

4.2 Class-aware Deviation Scaling

While ACC improves the MD in a feature-specific manner, it treats all classes uniformly when quantifying deviations from their prototypes, neglecting class-dependent characteristics. Specifically, classes with low intra-class variance may fail to adequately reflect subtle deviations, whereas classes with high variance are more susceptible to noise. Moreover, variations in class prototypes can further introduce bias in the resulting distance scores. To address this, we introduce **Class-aware Deviation Scaling (CDS)**, adjusting deviation features based on intra-class variance and prototype norms.

Variance-based scaling. For class c, the intra-class variance v_c reflects compactness in feature space. Tighter classes require amplified distance responses to detect subtle deviations, while looser classes need suppression to reduce noise sensitivity. We define the variance-based scaling factor:

$$s_v = \frac{\bar{v}}{v_c + \varepsilon}, \tag{9}$$

where $\bar{v} = \frac{1}{C} \sum_{c=1}^{C} v_c$ is the mean of variances across classes, and ε prevents division by zero.

Norm-based scaling. The norm $\|\mu_c\|$ reflects the overall strength of the class-level representation. Classes with higher norms correspond to stronger feature representations and are therefore amplified to emphasize their dominant directions, while classes with lower norms correspond to weaker representations and are suppressed to mitigate noise or imbalance across classes. We define the norm-based scaling factor:

$$s_n = \frac{\|\mu_c\|}{\frac{1}{C} \sum_{c=1}^{C} \|\mu_c\| + \varepsilon}, \tag{10}$$

where ε is a small constant to avoid division by zero.

The adjusted deviation feature is then defined as

$$w_d = d \cdot (\alpha_{\text{var}} \cdot s_v + \alpha_{\text{norm}} \cdot s_n), \tag{11}$$

where $d = f - f_a$ denotes the deviation feature of f from its class prototype f_a; α_{var} and α_{norm} are the weighting coefficients for the variance- and norm-based scaling factors.

The Mahalanobis distance is formulated as

$$d_{\hat{\mathcal{M}}}(f, f_a) = \sqrt{w_d^\top \hat{\mathcal{M}}(f) w_d}. \tag{12}$$

This formulation enables class-aware distance measurement without additional training, while preserving interpretability and stability.

4.3 Mahalanobis Gap Scoring

Distance-based OOD methods typically score a test sample by its nearest-prototype Mahalanobis distance, as in Eq. (12). However, this local criterion ignores the global distance structure and may fail when an OOD sample lies close to one ID prototype. Such hard samples often have similar distances to multiple prototypes, leading to flatter sorted distance profiles and smaller prototype gaps.

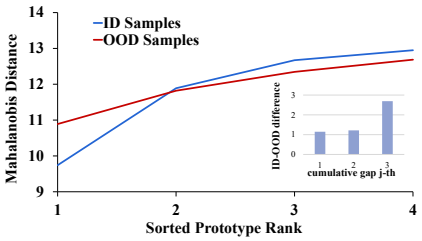


Figure 2: Sorted Mahalanobis distance profiles for ID and OOD samples. The inset shows ID-OOD differences: 1 for nearest-prototype scoring, and 2/3 for cumulative gaps up to the second/third prototypes.

Figure 2 illustrates this pattern using sorted Mahalanobis distance profiles, where each point denotes the 25th percentile of the j-th smallest distance z'_j computed separately over ID and OOD samples. This percentile emphasizes samples with smaller Mahalanobis distances, where hard OOD samples may appear close to ID prototypes. The flatter OOD profile and the inset comparison show that cumulative prototype gaps provide additional separation beyond nearest-prototype scoring. Motivated by this observation, we introduce Mahalanobis Gap Scoring (MGS), which exploits inter-class distance structure to improve scoring reliability in challenging nearest-prototype cases.

Let $f = h(x)$ denote the feature extracted from the sample x. The i-th class prototype among C classes is $f_a^{(i)}$. The Mahalanobis distance vector from f to all class prototypes is defined as

$$\mathbf{z} = [z_1, z_2, \ldots, z_C], \tag{13}$$

where $z_i = d_M(f, f_a^{(i)}), \quad i = 1, \ldots, C$. By sorting the elements of z in ascending order, we obtain:

$$\mathbf{z}' = [z'_1, z'_2, \ldots, z'_C], \tag{14}$$

where z'_j denotes the j-th smallest in $\mathbf{z}$, and z'_1 corresponds to the MD to the nearest prototype.

Then, the MGS is defined as

$$\mathrm{MGS}(x) = \frac{1}{C-1} \sum_{j=2}^{C} \left(z'_j - z'_1\right), \tag{15}$$

which reflects the separation of the sample from its nearest prototype compared to other classes.

Intuitively, a larger MGS indicates that the sample is not only close to its nearest prototype, but also well separated from other class prototypes, resulting in a more unambiguous neighborhood structure in the feature space. Consequently, a higher MGS corresponds to a higher confidence that the sample belongs to the ID.

To capture both relative and absolute distance information, we combine the MGS with the MD. MGS quantifies the relative separation of a sample from its nearest class prototypes with respect to other class prototypes, whereas the Mahalanobis distance captures its absolute proximity to the nearest prototype. This combination improves the reliability of distance-based discrimination.

The scoring function is defined as

$$S(x) = \lambda_{\text{gap}} \cdot \mathrm{MGS}(x) - d_M(f, f_a), \tag{16}$$

where $f = h(x)$ denotes the feature representation of the input sample x, and λ_{gap} balances the contributions of structural and absolute distance information. This scoring fully exploits the information contained in the Mahalanobis distance vector, rather than relying on a single element, thereby enhancing the reliability of the score.

Theorem 1. *Given a C-class ID classification task, let the Mahalanobis distance vector for a sample x be $z = [z_1, z_2, \ldots, z_C]$, where $z_c = d_{\widehat{M}}(f, \mu_c)$ denotes the calibrated Mahalanobis distance between the test feature $f = h(x)$ and the c-th class prototype μ_c. Let $z' = [z'_1, z'_2, \ldots, z'_C]$ denote the sorted distances in ascending order, such that $z'_1 = \min_c z_c$. Define the nearest-prototype score as $S_{\text{near}}(x) = -z'_1$, and the MGS-enhanced score as $S_\lambda(x) = \lambda_{\text{gap}}\mathrm{MGS}(x) - z'_1$. Let $U_{\text{FPR}} = \inf_{S_{\text{near}}(x) < \eta_{\text{near}},\, \mathrm{MGS}(x) > 0} \frac{\eta_\lambda + z'_1}{\mathrm{MGS}(x)}$. For threshold-compatible fixed thresholds η_{near} and η_λ, if $U_{\text{FPR}} > 0$ and $0 < \lambda_{\text{gap}} < U_{\text{FPR}}$, we have*

$$\mathrm{FPR}_{\mathrm{MGS}}(\lambda_{\text{gap}}, \eta_\lambda) \leq \mathrm{FPR}_{\text{near}}(\eta_{\text{near}}).$$

Table 1: Evaluation results of FPR@95↓ (%) and AUROC↑ (%) on standard benchmarks (ImageNet-1K as ID). **Bold numbers** indicate the best results, and underlined numbers indicate the second-best results.

Method	iNaturalist		Textures		OpenImage-O		ImageNet-O		SUN		Places		Mean results	
	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑
MSP	52.77	88.42	66.13	80.46	66.84	83.91	100.00	28.64	68.58	81.75	71.57	80.63	70.98	73.97
Energy	53.95	90.59	52.29	86.73	64.88	87.06	100.00	41.92	58.28	86.73	65.40	84.13	65.80	79.53
MahaVanilla	93.78	63.90	43.48	89.98	85.46	71.63	66.35	80.94	97.37	52.02	97.33	51.68	80.63	68.36
Mahalanobis++	**5.67**	**98.77**	66.52	83.91	76.42	78.66	**50.81**	**90.50**	65.20	82.56	**47.12**	**90.28**	51.96	87.45
KNN	59.55	86.02	10.85	97.49	64.29	82.89	62.40	84.62	69.64	80.71	78.03	74.86	57.46	84.43
VIM	71.33	87.53	14.79	96.86	59.01	88.81	85.05	70.87	82.05	81.10	83.27	74.42	65.92	83.93
maxLogit	50.87	91.13	54.33	86.40	63.90	87.36	100.00	40.86	59.90	86.59	65.68	84.18	65.78	79.42
KLMatch	44.29	89.91	64.77	82.82	61.48	85.74	100.00	38.67	73.51	80.55	75.56	78.86	69.93	76.09
ReAct	29.86	94.75	45.85	90.13	54.38	89.66	99.85	46.86	**45.28**	**90.51**	53.66	88.11	54.81	83.34
ODIN	50.86	91.14	54.31	86.40	63.88	87.36	100.00	40.86	59.90	86.59	65.67	84.18	65.77	79.42
FDBD	40.16	93.71	37.52	92.10	56.07	90.30	100.00	59.83	61.25	86.81	67.11	84.07	60.35	84.47
Neco	83.39	69.17	17.80	95.62	71.61	77.32	73.30	75.88	87.25	66.46	88.28	66.44	70.27	75.15
WDiscOOD	91.98	65.98	38.83	91.62	83.90	71.95	64.10	82.16	96.64	54.64	96.82	53.65	78.71	70.00
DCC	47.49	91.47	**5.66**	**98.80**	51.08	89.16	65.60	82.26	65.42	83.94	75.32	78.77	51.77	87.43
Ours	39.53	93.29	5.73	98.68	**40.72**	**92.31**	70.05	80.59	61.96	85.13	71.98	80.53	**48.33**	**88.42**

Table 2: Evaluation results of FPR@95↓ (%) and AUROC↑ (%) on standard benchmarks (CIFAR-100 as ID). **Bold numbers** indicate the best results, and underlined numbers indicate the second-best results.

Method	SVHN		Texture		LSUN_Resize		LSUN		iSUN		Places365		Mean results	
	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑
MSP	82.01	75.18	84.80	71.41	85.22	69.18	60.50	85.60	86.01	70.17	82.24	74.71	80.13	74.37
Energy	88.02	81.30	84.33	71.01	70.69	80.14	14.75	97.43	74.60	78.95	78.25	78.15	68.44	81.17
maxLogit	86.26	81.42	83.44	71.18	71.70	79.77	16.98	97.06	75.05	78.69	78.22	77.78	68.53	81.06
ODIN	86.24	81.43	83.42	71.19	71.71	79.77	16.99	97.06	75.03	78.69	77.77	**78.23**	68.53	81.06
MahaVanilla	54.26	88.01	28.32	92.15	38.33	92.08	96.47	44.43	36.09	92.38	95.04	56.29	58.08	77.56
Mahalanobis++	38.76	92.81	23.01	94.44	21.79	96.13	4.26	99.03	20.68	96.43	87.70	71.06	32.70	91.65
KNN	**17.84**	**96.35**	24.29	93.70	47.34	90.41	31.46	92.85	39.69	91.90	93.21	60.12	42.30	87.56
ReAct	83.69	82.77	79.79	77.70	71.56	81.71	15.82	97.13	75.19	81.06	78.25	77.53	67.38	82.98
Line	31.13	91.90	39.24	87.91	23.37	94.95	5.77	98.85	22.64	95.13	88.50	63.82	35.11	88.76
DICE	59.95	88.20	61.44	77.13	54.93	88.25	**0.91**	**99.74**	52.43	88.52	80.34	77.45	51.66	86.55
FDBD	49.75	91.38	43.30	91.59	57.75	88.71	19.19	96.54	56.03	89.43	**77.39**	77.87	50.57	89.25
DCC	22.93	95.56	19.77	95.69	21.73	96.13	6.05	98.78	21.06	96.38	87.52	71.87	29.84	92.40
Ours	19.12	96.15	**18.32**	**96.14**	**19.70**	**96.42**	5.04	98.92	**18.63**	**96.68**	83.76	74.91	**27.43**	**93.20**

Here, $\mathrm{FPR}_{\mathrm{MGS}}(\lambda_{\mathrm{gap}}, \eta_\lambda)$ *is the false positive rate based on* S_λ *with fixed threshold* η_λ*, and* $\mathrm{FPR}_{\mathrm{near}}(\eta_{\mathrm{near}})$ *is the false positive rate based on* S_{near} *with fixed threshold* η_{near}*.*

The proof is deferred to Appendix E.

5 Experiments

Datasets. We use **CIFAR-10** and **CIFAR-100** [20] as ID datasets, and evaluate on six standard OOD datasets: **SVHN** [33], **Textures (DTD)** [5], **Places365** [49], **LSUN** [47], **LSUN-Resize** [47], and **iSUN** [46]. To assess scalability under high-resolution inputs and large semantic spaces, we further use **ImageNet-1k** [6], containing over 1.2M training images from 1,000 categories, as the ID dataset. For ImageNet-scale OOD evaluation, we use **Textures (DTD)** [5], **SUN** [44], **Places** [49], **iNaturalist** [42], **ImageNet-O** [16], and **OpenImage-O** [19, 43]. All OOD datasets have non-overlapping categories with the corresponding ID dataset.

Baseline. We compare AMG with representative OOD detection methods, including confidence/logit-based methods (**MSP** [15], **Energy** [28], **MaxLogit** [40], **ODIN** [25]), distance-based methods (**Mahalanobis** [21], **Mahalanobis++** [32], **KNN** [40], **FDBD** [27], **DCC** [11]), feature/weight modification methods (**ReAct** [38], **DICE** [39], **LINE** [1]), and subspace-based methods (**ViM** [43], **Neco** [2], **WDiscOOD** [4]). All methods are evaluated with the same backbones and hardware settings, using hyperparameters from the original implementations. We report **AUROC** and **FPR@95**.

5.1 Main Result

CIFAR. As shown in Tables 2 and 3, the proposed method outperforms DCC and FDBD on most OOD datasets under the CIFAR-10/100 benchmarks with DenseNet [17]. On CIFAR-10, it achieves 11.19% FPR@95 and 97.77% AUROC, improving over DCC/FDBD by 3.44/2.67 points in FPR@95 and 0.94/0.54 points in AUROC. On CIFAR-100, it reduces FPR@95 by 2.41/23.14 points and improves AUROC by 0.80/3.95 points over DCC/FDBD. These results indicate that AMG provides a more reliable distance-based score by jointly adapting feature-space geometry and class-wise deviations, leading to improved OOD discrimination across diverse test distributions.

ImageNet-1k. As shown in Table 1, AMG demonstrates stable and competitive OOD detection performance on the large-scale and semantically complex ImageNet-1k benchmark with ResNet-50 [12]. Compared with DCC, it reduces the mean FPR@95 from 51.77% to 48.33% and improves

Table 3: Evaluation results of FPR@95↓ (%) and AUROC↑ (%) on standard benchmarks (CIFAR-10 as ID). **Bold numbers** indicate the best results, and underlined numbers indicate the second-best results.

Method	SVHN		Texture		LSUN_Resize		LSUN		iSUN		Places365		Mean results	
	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑
MSP	47.25	93.57	64.33	88.14	42.11	94.55	33.68	95.54	42.54	94.49	62.39	88.73	48.72	92.50
Energy	38.88	94.19	56.33	86.41	8.79	98.19	3.84	99.15	9.58	98.11	39.75	91.82	26.20	94.65
maxLogit	37.95	94.32	56.06	86.50	9.45	98.12	4.21	99.09	10.13	98.05	40.37	91.77	26.36	94.64
ODIN	37.83	94.33	56.05	86.52	9.48	98.12	4.21	99.09	10.15	98.05	40.37	91.77	26.35	94.65
MahaVanilla	7.93	98.11	25.66	92.81	51.71	87.91	67.99	82.62	45.52	88.79	87.02	65.15	47.64	85.90
Mahalanobis++	9.31	98.00	22.64	94.52	4.08	99.05	**0.32**	**99.85**	5.77	98.84	63.79	84.52	17.65	95.80
KNN	**3.96**	**99.29**	19.61	96.39	9.90	98.12	6.91	98.75	10.26	98.20	46.33	89.95	16.16	96.79
ReAct	33.40	95.23	48.00	91.29	7.51	98.40	4.63	99.06	8.44	98.27	39.54	92.30	23.59	95.76
Line	11.42	97.75	23.44	95.11	4.12	99.10	0.62	99.83	5.08	99.02	43.85	91.13	14.75	96.99
DICE	27.82	94.96	46.03	86.96	4.23	99.06	0.38	99.00	5.20	98.99	45.00	90.18	21.44	95.01
FDBD	6.20	98.61	23.07	95.94	5.56	98.82	3.49	99.32	5.85	98.77	38.97	91.94	13.86	97.23
DCC	8.21	98.41	16.51	96.41	3.01	**99.82**	0.71	99.82	4.04	99.17	55.28	87.84	14.63	96.83
Ours	7.58	98.66	**15.11**	**97.23**	**2.17**	99.50	2.21	99.56	**3.20**	**99.34**	**36.85**	**92.36**	**11.19**	**97.77**

the mean AUROC from 87.43% to 88.42%, while also outperforming ReAct and KNN by larger margins. On challenging OOD benchmarks such as iNaturalist and OpenImage-O, AMG further reduces FPR@95 by 7.96 and 15.35 points, respectively. These gains indicate that AMG better handles heterogeneous feature distributions and provides a more reliable distance-based score than nearest-prototype calibration or less discriminative distance computations. Overall, AMG achieves a favorable balance between low FPR@95 and high AUROC, enabling more reliable OOD detection under complex distributions.

5.2 Ablation Analysis

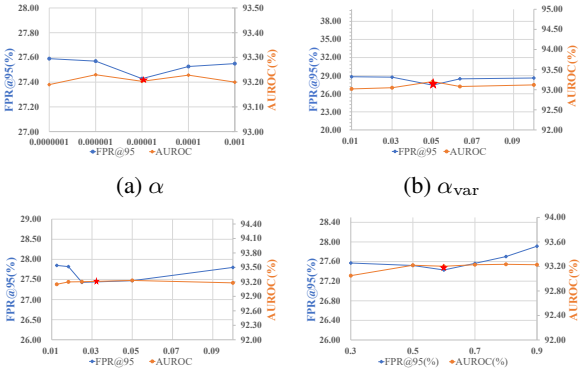
We conduct ablation experiments to evaluate each AMG component, as shown in Table 4. Removing ACC, CDS, and MGS increases FPR@95 by 2.46, 2.41, and 1.05 points, respectively, and decreases AUROC by 0.54, 0.38, and 0.70 points. These results show that adaptive covariance calibration, class-aware deviation scaling, and gap-based scoring each contribute to AMG by mitigating ID misclassification, handling uneven class distributions, and leveraging all prototype distances for uncertainty estimation. Overall, the full AMG model achieves the best performance, confirming the effectiveness and complementarity of its modules.

Table 4: Ablation study on different module.

Method	Mean Results	
	FPR@95 ↓	AUROC ↑
w/o MGS	28.48	92.50
w/o CDS	29.84	92.82
w/o ACC	29.87	92.66
Ours	**27.43**	**93.20**

5.3 Sensitivity Analysis

We evaluate the robustness and performance under variations of four semantically meaningful hyperparameters using FPR@95 and AUROC. Figure 3 presents the sensitivity of each parameter: the principal subspace parameter α, shown in Figure 3a, enhances discriminative stability; the variance-related parameter α_{var}, illustrated in Figure 3b, regulates the variance component; the prototype norm α_{norm}, shown in Figure 3c, controls the influence of prototype magnitude on deviation features; and λ_{gap}, illustrated in Figure 3d, weights the contribution of MGS to separate ID and OOD samples. These results demonstrate that appropriate hyperparameter configuration improves detection performance.



(a) α (b) α_{var} (c) α_{norm} (d) λ_{gap}

Figure 3: Parameter sensitivity analysis.

5.4 Generalization Analysis

As shown in Table 5, our method demonstrates consistent performance across three structurally distinct backbones: ResNet-50, Swin-B [29], and DeiT [41], indicating strong cross-backbone generalization. On ResNet-50, it achieves FPR@95 of 48.33% and AUROC of 88.42%, outperforming

Table 5: Performance of OOD Detection Methods on ImageNet-1k with ResNet-50, Swin-B, and DeiT Backbones. **Bold numbers** indicate the best results, and underlined numbers indicate the second-best results.

Method	ResNet-50		Swin-B		DeiT		Mean results	
	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑
MSP	70.98	73.97	63.49	81.29	66.97	79.80	67.15	78.35
Energy	65.80	79.53	60.35	80.07	72.65	71.65	66.27	77.08
MahaVanilla	80.63	68.36	52.07	87.86	73.86	83.98	68.85	80.07
KNN	57.46	84.43	65.24	85.08	76.24	82.75	66.31	84.09
VIM	65.92	83.93	**51.33**	86.77	71.13	83.91	62.79	84.87
maxLogit	65.78	79.42	59.56	80.94	66.38	76.43	63.91	78.93
KLMatch	69.93	76.09	63.85	81.83	67.24	82.68	67.01	80.20
ReAct	54.81	83.34	53.38	85.20	68.74	77.16	58.98	81.90
ODIN	54.77	79.42	58.94	80.68	66.43	76.07	60.05	78.72
FDB	60.35	84.47	55.75	86.57	71.84	82.78	62.65	84.61
Neco	70.27	75.15	54.89	81.73	**62.03**	79.20	62.40	78.69
WDiscOOD	78.71	70.00	–	–	73.83	83.97	76.27	76.99
DCC	51.77	87.43	52.07	88.18	68.20	84.93	57.35	86.85
Ours	**48.33**	**88.42**	51.98	**88.28**	65.63	**85.25**	**55.31**	**87.32**

DCC (51.77% / 87.43%) and reducing FPR@95 compared with ReAct and KNN by 6.48% and 9.13%. On Swin-B, our method attains FPR@95 of 51.98% and AUROC of 88.28%, matching or surpassing DCC and consistently outperforming ReAct and KNN. For DeiT, it achieves the second-lowest FPR@95 of 65.63% and the highest AUROC of 85.25%, again improving upon DCC. These results indicate that our approach remains effective across convolutional, hierarchical Transformer, and pure Transformer architectures, thanks to its test-time treatment of the global feature space, which stabilizes discrimination between ID and OOD samples.

5.5 Near-OOD Evaluation

Near-OOD detection is challenging because semantically related OOD samples may lie close to ID prototypes, making nearest-prototype distance unreliable. We evaluate this setting using CIFAR-10 and CIFAR-100 as mutually near-OOD benchmarks. As shown in Table 6, AMG improves over DCC in both directions: from 96.94/54.55 to 94.54/62.40 FPR@95/AUROC when CIFAR-100 is ID, and from 63.22/84.94 to 50.45/89.86 when CIFAR-10 is ID. These results show that global prototype-gap information improves detection under semantic overlap.

Table 6: Comparison of OOD detection performance on near-OOD scenarios.

Method	CIFAR-100 vs CIFAR-10		CIFAR-10 vs CIFAR-100	
	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑
DCC	96.94	54.55	63.22	84.94
Ours	**94.54**	**62.40**	**50.45**	**89.86**

5.6 Covariate-Shift Evaluation

To evaluate generalization, we assess our method on distributional shifts where class labels are unchanged, using CIFAR-10 and CIFAR-100 as ID datasets and CIFAR10C and CIFAR100C as shifted samples [14]. As shown in Table 7, our method outperforms DCC, reducing FPR@95 and increasing AUROC on both datasets. On CIFAR10C, FPR@95 decreases from 56.11% to 51.93% and AUROC rises from 82.95% to 85.01%, while on CIFAR100C, FPR@95 drops from 59.24% to 54.56% and AUROC improves from 82.97% to 85.05%, demonstrating stronger robustness under distributional shifts.

Table 7: Performance under distributional shifts. CIFAR-10/100 serve as ID datasets, while their corrupted variants, CIFAR10C and CIFAR100C, are used as shifted test samples.

Method	CIFAR-10 vs CIFAR10C		CIFAR-100 vs CIFAR100C	
	FPR@95 ↓	AUROC ↑	FPR@95 ↓	AUROC ↑
DCC	56.11	82.95	59.24	82.97
Ours	**51.93**	**85.01**	**54.56**	**85.05**

5.7 Generalizability of MGS

To evaluate the generalizability of MGS, we further integrate it into two representative distance-based OOD detection methods, Mahalanobis++ and DCC, on CIFAR-10. For each baseline, we only replace the original nearest-prototype scoring rule with the MGS-enhanced score, while keeping all other settings unchanged. As shown in Table 8, incorporating MGS consistently im-

Table 8: Generalizability of MGS on CIFAR-10.

Method	Scoring	FPR@95↓	AUROC↑
Mahalanobis++	Original	17.65	95.80
	+ MGS	15.76	96.62
DCC	Original	14.63	96.83
	+ MGS	12.13	97.58
AMG (Ours)		**11.19**	**97.77**

proves the performance of both methods. Specifically, MGS reduces the mean FPR@95 of Mahalanobis++ from 17.65% to 15.76% and increases AUROC from 95.80% to 96.62%. Similarly, for DCC, MGS reduces FPR@95 from 14.63% to 12.13% and improves AUROC from 96.83% to 97.58%. These results demonstrate that MGS is not restricted to the proposed AMG framework but can serve as a generalizable scoring mechanism for distance-based OOD detection methods.

5.8 Robustness under Limited ID Data

To evaluate the robustness of AMG under limited ID data, we conduct experiments on CIFAR-10 using only 10% of the original ID training samples as reference data, while keeping the backbone and evaluation protocol unchanged. As shown in Table 9, AMG achieves a mean FPR@95 of 14.58% and AUROC of 97.10%, outperforming DCC by 2.02 and 0.85 percentage points, respectively. Compared with Mahalanobis++, AMG further improves FPR@95 and AUROC by 4.67 and 1.73 percentage points, respectively. These results demonstrate that AMG maintains strong OOD detection performance even with limited ID reference data, benefiting from adaptive covariance calibration and global class-wise distance information.

Table 9: OOD detection performance on CIFAR-10 using only 10% of ID training samples.

OOD Dataset	Mahalanobis++		DCC		Ours	
	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑	FPR@95↓	AUROC↑
SVHN	12.57	97.34	**11.68**	97.58	12.11	**97.84**
Texture	24.79	93.74	17.54	96.01	**16.28**	**96.93**
LSUN-Resize	5.73	98.82	4.71	99.00	**4.34**	**99.14**
LSUN	**0.36**	**99.85**	0.54	99.81	2.03	99.58
iSUN	6.99	98.61	6.39	98.75	**6.13**	**98.83**
Places365	65.05	83.88	58.77	86.35	**46.57**	**90.27**
Mean	19.25	95.37	16.60	96.25	**14.58**	**97.10**

5.9 Computational Efficiency

We evaluate the computational efficiency of AMG by comparing its runtime and OOD detection performance with DCC under identical experimental settings. As shown in Table 10, the DCC-style implementation of AMG achieves better detection performance than DCC, but incurs a 17% increase in runtime. In contrast, the optimized implementation reduces the runtime from 11 min 59 s to 8 min 30 s, achieving a 29.1% reduction while maintaining nearly identical detection performance. Notably, the optimized AMG is 17.3% faster than DCC while achieving superior FPR@95 and AUROC. These results demonstrate that AMG effectively balances detection performance and computational efficiency. Further implementation details and complexity analysis are provided in Appendix H.

Table 10: Runtime and OOD detection performance on CIFAR-100.

Method	Runtime ↓	Relative ↓	FPR@95 ↓	AUROC ↑
DCC	10m 17s	1.00×	29.84	92.40
AMG (DCC-style)	11m 59s	1.17×	27.43	**93.20**
AMG (Optimized)	**8m 30s**	0.83×	**27.40**	**93.20**

6 Conclusion

In this work, we proposed Adaptive Mahalanobis Gap (AMG), a structure-aware distance-based OOD detection method. AMG adaptively reinforced the principal subspace at inference, reducing ID misclassification under distorted geometries. Class-wise variance and prototype norm refined deviation features, while Mahalanobis Gap Scoring leveraged Mahalanobis Distance information to improve uncertainty estimation. Experiments on CIFAR-10/100 and ImageNet-1k showed that AMG achieves competitive or superior performance, with stronger mean FPR@95 and AUROC over strong OOD detection baselines.

Acknowledgements and Disclosures

We acknowledge financial support from the Shandong Provincial Project for Enhancing the Innovation Capacity of Science and Technology Small and Medium-sized Enterprises (No. 2023TSGC0115). The authors declare no competing interests.

References

[1] Yong Hyun Ahn, Gyeong-Moon Park, and Seong Tae Kim. 2023. Line: Out-of-distribution detection by leveraging important neurons. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19852–19862. IEEE.

[2] Mouin Ben Ammar, Nacim Belkhir, Sebastien Popescu, Antoine Manzanera, and Gianni Franchi. 2024. Neco: Neural collapse based out-of-distribution detection. In *The Twelfth International Conference on Learning Representations*.

[3] Sima Behpour, Thang Long Doan, Xin Li, Wenbin He, Liang Gou, and Liu Ren. 2023. Gradorth: A simple yet efficient out-of-distribution detection with orthogonal projection of gradients. *Advances in Neural Information Processing Systems*, 36:38206–38230.

[4] Yiye Chen, Yunzhi Lin, Ruinian Xu, and Patricio A Vela. 2023. Wdiscood: Out-of-distribution detection via whitened linear discriminant analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5298–5307.

[5] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613.

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

[7] Mostafa ElAraby, Sabyasachi Sahoo, Yann Pequignot, Paul Novello, and Liam Paull. 2023. Grood: Gradient-aware out-of-distribution detection in interpolated manifolds. *arXiv preprint arXiv:2312.14427*.

[8] Angelos Filos, Panagiotis Tigkas, Rowan McAllister, Nicholas Rhinehart, Sergey Levine, and Yarin Gal. 2020. Can autonomous vehicles identify, recover from, and adapt to distribution shifts? In *International Conference on Machine Learning*, pages 3145–3153. PMLR.

[9] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. 2021. Exploring the limits of out-of-distribution detection. *Advances in neural information processing systems*, 34:7068–7081.

[10] Wei Fu, Yufei Chen, Wei Liu, Xiaodong Yue, and Chao Ma. 2023. Evidence reconciled neural network for out-of-distribution detection in medical images. In *International conference on medical image computing and computer-assisted intervention*, pages 305–315. Springer.

[11] Kaiyu Guo, Zijian Wang, Tan Pan, Brian C Lovell, and Mahsa Baktashmotlagh. 2025. Improving out-of-distribution detection via dynamic covariance calibration. *arXiv preprint arXiv:2506.09399*.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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

[14] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

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

[16] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271.

[17] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708.

[18] Rui Huang, Andrew Geng, and Yixuan Li. 2021. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689.

[19] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, and 1 others. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18.

[20] Alex Krizhevsky, Geoffrey Hinton, and 1 others. 2009. Learning multiple layers of features from tiny images.

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

[22] Wenhao Li, Xianjing Meng, Qiangchang Wang, Zhongyi Han, Zhibin Wu, and Yilong Yin. 2026. Dvla-rl: Dual-level vision-language alignment with reinforcement learning gating for few-shot learning. In *ICLR*.

[23] Wenhao Li, Qiangchang Wang, Xianjing Meng, Zhibin Wu, and Yilong Yin. 2025. Vt-fsl: Bridging vision and text with llms for few-shot learning. In *NeurIPS*, volume 38, pages 113697–113727.

[24] Jian Liang, Ran He, and Tieniu Tan. 2025. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64.

[25] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

[26] Jiaqi Liu, Guoyang Xie, Jinbao Wang, Shangnian Li, Chengjie Wang, Feng Zheng, and Yaochu Jin. 2024. Deep industrial image anomaly detection: A survey. *Machine Intelligence Research*, 21(1):104–135.

[27] Litian Liu and Yao Qin. 2023. Fast decision boundary based out-of-distribution detector. *arXiv preprint arXiv:2312.11536*.

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

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

[30] Shuo Lu, Yingsheng Wang, Lijun Sheng, Lingxiao He, Aihua Zheng, and Jian Liang. 2025. Out-of-distribution detection: A task-oriented survey of recent advances. *ACM Computing Surveys*, 58(2):1–39.

[31] Peyman Morteza and Yixuan Li. 2022. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7831–7840.

[32] Maximilian Mueller and Matthias Hein. 2025. Mahalanobis++: Improving ood detection via feature normalization. *arXiv preprint arXiv:2505.18032*.

[33] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, and 1 others. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 7. Granada.

[34] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436.

[35] Bo Peng, Yadan Luo, Yonggang Zhang, Yixuan Li, and Zhen Fang. 2024. Conjnorm: Tractable density estimation for out-of-distribution detection. *arXiv preprint arXiv:2402.17888*.

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

[37] Vikash Sehwag, Mung Chiang, and Prateek Mittal. 2021. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*.

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

[39] Yiyou Sun and Yixuan Li. 2022. Dice: Leveraging sparsification for out-of-distribution detection. In *European conference on computer vision*, pages 691–708. Springer.

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

[41] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR.

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

[43] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. 2022. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4921–4930.

[44] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE.

[45] Kai Xu, Rongyu Chen, Gianni Franchi, and Angela Yao. 2023. Scaling for training time and post-hoc out-of-distribution detection enhancement. *arXiv preprint arXiv:2310.00227*.

[46] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. 2015. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*.

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

[48] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*.

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

A Limitations

One limitation is that we mainly consider image classification tasks with frozen visual encoders, and leave extending AMG to broader test-time OOD detection scenarios, such as open-vocabulary recognition, dense prediction, multimodal understanding, and streaming environments, for future exploration.

B Implementation details.

We run all OOD detection experiments on NVIDIA GeForce RTX-4090Ti GPUs with PyTorch 2.8.0. For conventional OOD detection on CIFAR benchmarks, we use DenseNet and WRN backbones pre-trained on the corresponding ID training sets. For ImageNet-scale experiments, we use a pre-trained ResNet-50 backbone.

AMG contains four main hyperparameters: α, α_{var}, α_{norm}, and λ_{gap}, which control covariance calibration, variance-based scaling, norm-based scaling, and gap-based scoring, respectively. Hyperparameters are selected on the validation split for each evaluation setting and then kept fixed during testing. The default values are $\alpha = 10^{-5}$, $\alpha_{\mathrm{var}} = 10$, $\alpha_{\mathrm{norm}} = 300$, and $\lambda_{\mathrm{gap}} = 0.6$ unless otherwise specified.

C Broader Impacts

This work may have positive societal impacts by improving the reliability of machine learning systems in safety-sensitive applications, such as autonomous driving, medical image analysis, and industrial inspection, where detecting out-of-distribution inputs is important. However, potential negative impacts may arise if the detector is over-relied upon or deployed without sufficient validation. False positives may reject valid inputs, while false negatives may fail to identify unsafe OOD samples. Therefore, AMG should be carefully evaluated under the target deployment conditions before being used in real-world systems.

D Declaration of LLM usage

This work does not use LLMs as an important, original, or non-standard component of the proposed method, experimental design, data generation, or evaluation. Any use of LLMs, if applicable, was limited to writing, editing, or formatting assistance and did not affect the core methodology, scientific rigor, or originality of the research.

E Theoretical Analysis of Mahalanobis Gap Scoring

We provide the proof of Theorem 1. Let Q_{out} denote the OOD distribution. For fixed thresholds η_{near} and η_λ, the false positive rates of nearest-prototype scoring and MGS-enhanced scoring are

$$\mathrm{FPR}_{\mathrm{near}}(\eta_{\mathrm{near}}) = Q_{\mathrm{out}}[S_{\mathrm{near}}(x) \geq \eta_{\mathrm{near}}]\,,$$

and

$$\mathrm{FPR}_{\mathrm{MGS}}(\lambda_{\mathrm{gap}}, \eta_\lambda) = Q_{\mathrm{out}}[S_\lambda(x) \geq \eta_\lambda]\,.$$

By definition,

$$S_{\mathrm{near}}(x) = -z_1', \qquad S_\lambda(x) = \lambda_{\mathrm{gap}}\mathrm{MGS}(x) - z_1'.$$

We call the thresholds η_{near} and η_λ threshold-compatible if every sample with zero Mahalanobis gap that is rejected by nearest-prototype scoring is also rejected by MGS-enhanced scoring:

$$S_{\mathrm{near}}(x) < \eta_{\mathrm{near}},\ \mathrm{MGS}(x) = 0 \Rightarrow S_\lambda(x) < \eta_\lambda.$$

Since $S_\lambda(x) = S_{\mathrm{near}}(x)$ when $\mathrm{MGS}(x) = 0$, a sufficient condition for threshold compatibility is $\eta_\lambda \geq \eta_{\mathrm{near}}$.

We now show that, under the safe range of λ_{gap}, the false-positive set of S_λ is contained in that of S_{near}. Consider an OOD sample rejected by nearest-prototype scoring:

$$S_{\mathrm{near}}(x) < \eta_{\mathrm{near}}.$$

If $\mathrm{MGS}(x) = 0$, then the sample is also rejected by S_λ due to threshold compatibility. If $\mathrm{MGS}(x) > 0$, requiring the sample to be rejected by S_λ gives

$$S_\lambda(x) < \eta_\lambda,$$

which is equivalent to

$$\lambda_{\mathrm{gap}}\mathrm{MGS}(x) - z_1' < \eta_\lambda.$$

Rearranging yields

$$\lambda_{\text{gap}} < \frac{\eta_\lambda + z_1'}{\text{MGS}(x)}.$$

Therefore, if

$$0 < \lambda_{\text{gap}} < U_{\text{FPR}} = \inf_{S_{\text{near}}(x) < \eta_{\text{near}},\, \text{MGS}(x) > 0} \frac{\eta_\lambda + z_1'}{\text{MGS}(x)},$$

then every OOD sample rejected by S_{near} is also rejected by S_λ.

Thus,

$$\{x : S_\lambda(x) \geq \eta_\lambda\} \subseteq \{x : S_{\text{near}}(x) \geq \eta_{\text{near}}\}.$$

Taking probability over $x \sim Q_{\text{out}}$ gives

$$Q_{\text{out}}[S_\lambda(x) \geq \eta_\lambda] \leq Q_{\text{out}}[S_{\text{near}}(x) \geq \eta_{\text{near}}]\,.$$

Hence,

$$\text{FPR}_{\text{MGS}}(\lambda_{\text{gap}}, \eta_\lambda) \leq \text{FPR}_{\text{near}}(\eta_{\text{near}}).$$

This completes the proof.

If the set $\{x : S_{\text{near}}(x) < \eta_{\text{near}},\, \text{MGS}(x) > 0\}$ is empty, we set $U_{\text{FPR}} = +\infty$.

F Algorithm Box

Algorithm 1: Adaptive Mahalanobis Gap Scoring

Input: test sample x, encoder h, prototypes $\{\mu_c\}_{c=1}^C$, covariance Σ, bases A, B;
scaling factors $\{s_v(c), s_n(c)\}_{c=1}^C$, hyperparameters $\alpha, \beta, \alpha_{\text{var}}, \alpha_{\text{norm}}, \lambda_{\text{gap}}$
Output: OOD score $S(x)$

// Stage 1: Adaptive covariance calibration

$f \leftarrow h(x)$

foreach *subspace* $T \in \{A, B\}$ **do**

  $u_T \leftarrow Tf$

  $f_T \leftarrow T^\top u_T$

// Calibrate covariance and precision

$\widehat{\Sigma}(f) \leftarrow \Sigma + \alpha f_A f_A^\top - \beta f_B f_B^\top, \quad \widehat{M}(f) \leftarrow \widehat{\Sigma}(f)^{-1}$

// Stage 2: Class-aware Mahalanobis distances

for $c \leftarrow 1$ **to** C **do**

  $\gamma_c \leftarrow \alpha_{\text{var}} s_v(c) + \alpha_{\text{norm}} s_n(c)$

  $w_c \leftarrow \gamma_c (f - \mu_c)$

  $z_c \leftarrow \sqrt{w_c^\top \widehat{M}(f) w_c}$

// Stage 3: Mahalanobis gap scoring

Sort $z = [z_1, \ldots, z_C]$ in ascending order as $z' = [z_1', \ldots, z_C']$

$G(x) \leftarrow \frac{1}{C-1} \sum_{j=2}^{C} (z_j' - z_1')$

// Stage 4: Final AMG score

$S(x) \leftarrow \lambda_{\text{gap}} G(x) - z_1'$

Return: $S(x)$

G Additional Experiments

We use Wide ResNet (WRN) as the backbone network for CIFAR-100, as summarized in Table 11.

Additional results on CIFAR-100. Table 11 reports additional OOD detection results with CIFAR-100 as the ID dataset. Overall, AMG achieves the best mean performance among all compared methods, obtaining the lowest average FPR@95 of 43.29% and the highest average AUROC of 89.59%. Compared with the strongest Mahalanobis-based baseline DCC, AMG reduces the mean FPR@95 from 44.89% to 43.29% and improves the mean AUROC from 89.08% to 89.59%, demonstrating the benefit of combining adaptive covariance calibration, class-aware deviation scaling, and gap-based scoring.

Table 11: Additional OOD detection results with CIFAR-100 as ID. FPR@95↓ and AUROC↑ are reported. Bold indicates the best result, and underline indicates the second-best result.

Method	SVHN		Texture		LSUN-R		LSUN		iSUN		Places365		Mean	
	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC
MSP	85.24	70.91	85.89	71.39	79.82	79.62	49.10	88.56	81.57	78.10	83.36	73.41	77.48	76.93
Energy	87.59	70.87	85.27	72.85	78.49	82.20	16.74	96.83	82.51	80.89	**79.83**	74.94	79.83	71.95
MahaVanilla	44.91	89.68	42.38	90.03	**28.28**	91.73	99.71	47.63	40.56	91.16	91.94	65.66	59.63	79.35
KNN	43.03	89.95	40.99	89.74	39.49	**94.32**	56.30	82.02	34.47	92.32	85.65	69.68	49.82	86.34
MaxLogit	86.59	71.22	85.07	73.07	78.18	82.78	21.50	96.19	81.89	81.14	80.99	75.13	72.37	79.92
ReAct	87.63	74.61	84.33	76.22	79.76	81.27	17.10	96.77	83.53	79.92	81.22	**75.36**	72.62	80.69
ODIN	86.58	71.23	85.09	73.07	78.20	82.70	21.51	96.29	81.93	81.15	80.99	75.13	72.38	79.93
GEM	43.98	90.42	39.69	90.85	33.00	93.69	98.91	60.19	36.98	92.31	90.34	68.68	57.15	82.71
DICE	90.41	68.38	79.20	74.10	81.80	78.80	**4.18**	**98.68**	86.64	78.20	83.37	74.16	71.04	78.44
LINE	94.43	68.01	79.59	70.86	97.83	56.92	34.63	92.64	98.33	58.86	95.88	53.08	82.55	66.33
FDBD	73.03	83.18	71.97	83.26	58.07	87.45	45.85	91.16	63.22	88.18	82.51	75.31	65.77	85.14
DCC	36.07	92.64	**34.66**	**92.55**	32.29	94.04	50.83	87.75	33.48	93.14	82.00	74.35	44.89	89.08
Ours	**30.71**	**93.65**	34.75	92.41	34.17	93.79	45.96	89.11	**33.41**	**93.23**	80.71	75.35	**43.29**	**89.59**

Table 12: Evaluation results of FPR@95↓ (%) and AUROC↑ (%) on standard benchmarks with ImageNet-1k as ID and DeiT as backbone. Bold numbers indicate the best results, and underlined numbers indicate the second-best results.

Method	Texture		SUN		Places		iNaturalist		ImageNet-O		OpenImage-O		Mean results	
	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC
MSP	64.56	81.59	68.35	80.72	70.04	80.37	61.18	88.46	83.00	63.24	60.71	84.39	66.97	79.85
Energy	65.37	77.50	75.96	70.43	77.48	69.17	67.14	78.02	87.70	60.12	66.63	74.55	72.65	71.95
MahaVanilla	78.33	82.78	78.53	81.93	77.29	80.88	73.96	85.64	90.55	77.56	62.83	89.73	73.86	83.98
KNN	59.91	86.57	84.68	78.98	84.10	76.97	55.92	**98.64**	87.10	76.26	67.70	87.80	76.24	82.75
VIM	75.96	83.30	74.79	81.93	74.20	80.06	49.94	92.90	89.70	75.43	62.22	89.84	71.33	76.31
maxLogit	61.38	80.26	68.45	76.12	70.18	**85.44**	53.28	85.23	84.70	61.04	60.31	80.52	66.18	83.93
KLM	63.37	84.14	72.91	**82.34**	74.76	81.56	49.64	90.46	83.50	69.71	62.90	87.56	67.43	82.68
KLMatch	61.48	80.42	70.65	77.05	72.26	75.86	61.11	84.38	83.50	64.41	60.26	80.86	68.24	77.16
ReAct	64.77	80.41	76.42	75.63	70.76	74.94	52.92	84.14	84.60	60.54	60.47	89.15	67.34	76.07
ODIN	70.09	82.66	74.80	81.18	74.58	80.00	62.74	89.47	86.50	75.16	62.32	88.22	71.84	82.78
FDB	60.92	79.65	**65.62**	77.48	**68.02**	76.79	38.74	91.59	84.20	63.72	53.95	85.96	72.83	78.20
WDiscOOD	78.24	82.91	78.50	81.93	77.27	80.87	55.59	92.44	90.60	76.25	62.78	89.74	73.83	83.97
DCC	**58.42**	**87.13**	89.88	79.50	89.45	78.17	35.70	94.54	**80.20**	**78.89**	55.55	91.34	68.20	84.93
Ours	61.03	86.76	88.17	80.02	86.82	78.94	**23.42**	95.67	82.20	78.49	**52.14**	**91.64**	**65.63**	**85.25**

The improvement is particularly clear on SVHN, where AMG reduces FPR@95 by 5.36 percentage points and improves AUROC by 1.01 percentage points over DCC. On LSUN, AMG also achieves a lower FPR@95 and higher AUROC than DCC, indicating that the proposed scoring strategy is more effective when OOD samples present ambiguous nearest-prototype assignments. For iSUN and Places365, AMG remains consistently competitive and slightly improves over DCC on both FPR@95 and AUROC.

Although AMG does not achieve the best result on every individual OOD dataset, such as Texture and LSUN-Resize, it provides the best overall trade-off across all benchmarks. These results suggest that AMG is not tailored to a specific OOD distribution but instead improves the robustness of Mahalanobis-based detection across heterogeneous OOD scenarios. The consistent gain in the mean results further supports the complementarity of ACC, CDS, and MGS in producing a more reliable distance-based OOD score.

Additional results on ImageNet-1k. As shown in Table 12, AMG achieves the best mean performance on ImageNet-1k with the DeiT backbone, obtaining 65.63% FPR@95 and 85.25% AUROC. Compared with DCC, AMG reduces the mean FPR@95 by 2.57 percentage points and improves the mean AUROC by 0.32 percentage points. The improvement is especially pronounced on iNaturalist and OpenImage-O, where AMG achieves the lowest FPR@95 of 23.42% and 52.14%, respectively, together with strong AUROC performance. These results suggest that AMG remains effective on pure Transformer features, indicating that its adaptive covariance calibration, class-aware deviation scaling, and gap-based scoring are not limited to convolutional backbones. Although DCC performs better on Texture and ImageNet-O, AMG provides a stronger overall trade-off across heterogeneous OOD benchmarks.

As shown in Table 13, AMG achieves competitive performance on ImageNet-1k with the Swin-B backbone. Compared with DCC, AMG slightly reduces the mean FPR@95 from 52.07% to 51.98% and improves the mean AUROC from 88.18% to 88.28%. AMG obtains the best mean AUROC and the second-best mean FPR@95, indicating a favorable overall trade-off across diverse OOD benchmarks. The improvement is most visible on OpenImage-O, where AMG achieves the lowest FPR@95 of 31.33% while maintaining comparable AUROC. Although DCC remains stronger on

Table 13: OOD detection results with ImageNet-1k as ID and Swin-B as the backbone. FPR@95↓ and AUROC↑ are reported. Bold indicates the best result, and underline indicates the second-best result.

Method	Texture		SUN		Places		iNaturalist		ImageNet-O		OpenImage-O		Mean results	
	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC	FPR@95	AUROC
MSP	61.81	82.84	63.29	83.90	64.57	83.67	44.00	89.92	89.40	62.14	57.87	85.25	63.49	81.29
Energy	53.37	83.88	62.43	81.03	63.50	80.18	43.54	88.16	83.30	65.04	55.97	82.14	60.35	80.07
MahaVanilla	47.07	88.97	60.65	86.78	64.62	85.53	21.05	95.67	83.45	77.13	35.58	89.47	52.07	87.86
KNN	48.51	88.73	74.95	83.20	76.69	81.44	51.95	91.32	84.45	76.32	54.93	93.47	65.24	85.08
MaxLogit	45.02	88.26	60.17	84.65	64.07	82.04	22.42	95.93	81.15	77.89	36.24	93.29	**51.33**	86.77
ReAct	56.10	83.72	59.06	82.35	61.54	80.88	39.77	89.69	86.20	63.54	53.57	84.00	59.56	80.94
KLM	60.28	83.23	69.05	83.49	71.28	82.61	42.61	90.44	84.65	64.64	55.22	86.54	63.55	81.83
ReActMatch	50.83	87.55	54.38	**86.89**	57.20	85.87	31.27	93.31	82.80	68.67	44.70	89.13	53.85	85.20
ODIN	55.66	83.46	59.32	82.48	60.90	81.83	38.65	89.72	85.95	62.97	53.14	93.61	58.94	80.68
FDBD	50.69	87.44	59.78	86.84	62.58	**85.89**	33.61	93.71	84.85	74.27	43.01	81.26	55.75	86.57
Neco	54.50	82.65	**53.65**	84.17	**56.32**	83.35	24.96	93.82	92.65	60.66	46.36	85.43	54.74	81.73
DCC	**41.70**	**90.13**	71.47	85.12	74.99	83.90	**14.03**	**96.78**	**78.85**	**78.90**	31.37	**94.23**	52.07	88.18
Ours	42.89	89.97	70.61	85.61	73.64	84.46	14.14	96.75	79.30	78.69	**31.33**	94.21	51.98	**88.28**

Texture, iNaturalist, and ImageNet-O, AMG preserves competitive performance across most datasets, suggesting that the proposed scoring strategy generalizes to hierarchical Transformer features.

H Computational Efficiency of AMG

We provide two implementations of AMG to evaluate its computational efficiency. The first follows the direct matrix-inversion procedure adopted in the original DCC implementation, while the second employs the Woodbury identity to accelerate covariance inversion.

Direct Matrix Inversion. For each test feature f, let $f_A, f_B \in \mathbb{R}^D$ denote its principal- and residual-subspace components, respectively. AMG constructs the sample-dependent covariance matrix as

$$\Sigma'(f) = \Sigma + \alpha f_A f_A^\top - \beta f_B f_B^\top, \tag{17}$$

where $\Sigma \in \mathbb{R}^{D\times D}$ is the base covariance matrix, and $\alpha, \beta > 0$ are the calibration coefficients.

The DCC-style implementation directly computes the inverse covariance matrix for each test sample:

$$M(f) = \left(\Sigma + \alpha f_A f_A^\top - \beta f_B f_B^\top\right)^{-1}. \tag{18}$$

This requires a full $D \times D$ matrix inversion per sample, resulting in an approximate computational complexity of $O(D^3)$ for the inversion operation.

Woodbury-Based Optimization. To reduce the computational overhead, we further develop a mathematically equivalent implementation based on the Woodbury identity. Since the covariance calibration involves only two rank-one updates, we define

$$U(f) = [f_A, f_B] \in \mathbb{R}^{D\times 2}, \qquad R = \mathrm{diag}(\alpha, -\beta). \tag{19}$$

Accordingly, Eq. (17) can be rewritten as

$$\Sigma'(f) = \Sigma + U(f) R U(f)^\top. \tag{20}$$

Using the Woodbury identity, the inverse covariance matrix can be expressed as

$$\begin{aligned} M(f) = \Sigma^{-1} &- \Sigma^{-1} U(f) \\ &\cdot \left(R^{-1} + U(f)^\top \Sigma^{-1} U(f)\right)^{-1} U(f)^\top \Sigma^{-1}. \end{aligned} \tag{21}$$

Since $U(f)$ contains only two update directions, the inner matrix in Eq. (21) has dimension 2×2. By precomputing Σ^{-1} once, the optimized implementation replaces the per-sample $D \times D$ matrix inversion with a low-rank update. Consequently, the inversion-related complexity is reduced from $O(D^3)$ per sample to approximately $O(D^2)$ per sample, after a one-time $O(D^3)$ precomputation.

Table 14: Runtime and OOD detection performance on CIFAR-100. All methods are evaluated under the same hardware and experimental settings.

Method	Runtime ↓	Relative ↓	FPR@95 ↓	AUROC ↑
DCC	10m 17s	$1.00\times$	29.84	92.40
AMG (DCC-style)	11m 59s	$1.17\times$	27.43	**93.20**
AMG (Optimized)	**8m 30s**	$0.83\times$	**27.40**	**93.20**

Runtime Comparison. As shown in Table 14, the DCC-style implementation of AMG requires 11 min 59 s, corresponding to $1.17\times$ the runtime of DCC. In contrast, the optimized implementation reduces the runtime to 8 min 30 s, achieving a 29.1% reduction compared with the DCC-style implementation and a 17.3% reduction compared with DCC.

Importantly, the optimization only changes the numerical procedure for computing the inverse covariance matrix, without modifying ACC, CDS, MGS, or the final OOD scoring function. Therefore, both implementations are mathematically equivalent, subject to the required matrix invertibility conditions, with only minor differences arising from floating-point computations. The AUROC remains unchanged at 93.20%, while FPR@95 changes marginally from 27.43% to 27.40%.

These results demonstrate that AMG achieves improved OOD detection performance while maintaining competitive computational efficiency.

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: Yes, the experiments in Section 5 back up our claims.

   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: Yes, please see the Appendix A

   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: Yes, please see the Appendix E

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: Yes, please see the Appendix B

   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 adopted datasets are public datasets. The code is available.

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: Yes, please see the Appendix B

   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: [N/A]

   Justification: Our method is a post-hoc OOD detection approach built on fixed pre-trained models, without stochastic training or model updating. Therefore, the reported results are stable under the same backbone, dataset split, and hyperparameter configuration.

   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: Yes, please see the Appendix B

   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: Yes, research conducted in the paper conform 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: Please see the Appendix C

    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: This paper does not have a misuse risk.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The adopted datasets and compared baselines are properly credited.

    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: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor 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: The paper does not involve crowdsourcing nor 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: [Yes]

Justification: Please see the Appendix D

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.