Adaptive Covariance and Multi-Layer Alignment for Out-of-Distribution Detection

Haoyang Su
Australian Institute for Machine Learning
Adelaide University
Adelaide, SA 5005
haoyang.su@adelaide.edu.au

Qi Chen
Australian Institute for Machine Learning
Adelaide University
Adelaide, SA 5005
qi.chen04@adelaide.edu.au

Max Gutbrod
Regensburg Medical Image Computing
OTH Regensburg
Regensburg, BY 93053
max.gutbrod@oth-regensburg.de

Johan Verjans
Australian Institute for Machine Learning
Adelaide University
Adelaide, SA 5005
johan.verjans@adelaide.edu.au

Zhibin Liao *
Australian Institute for Machine Learning
Adelaide University
Adelaide, SA 5005
zhibin.liao@adelaide.edu.auu

Abstract

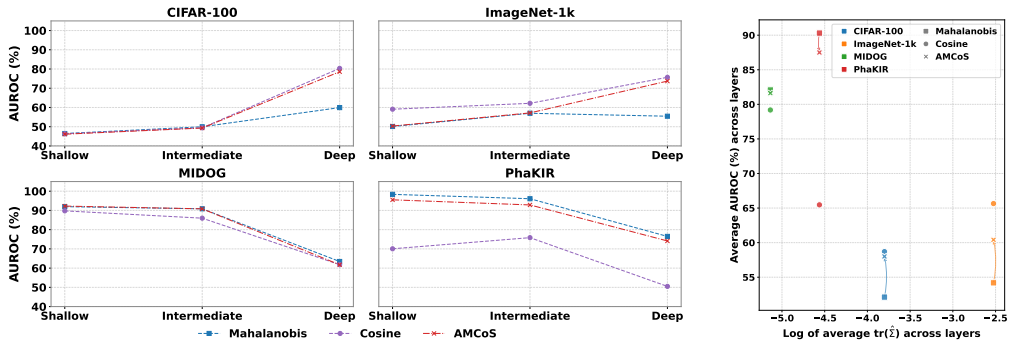
Out-of-distribution (OOD) detection plays a pivotal role in ensuring the reliability and trustworthiness of AI systems. Although existing approaches achieve strong performance by leveraging features, logits, or both, most of them cannot generalize well across domains due to either overlooking critical information in low-level representations from shallow layers or utilizing sub-optimal layer selection strategies. In this paper, we propose **M**ulti-Layer **A**daptive **M**ahalanobis-**Co**sine **S**imilarity (**ML-AMCoS**), a method that adaptively integrates cosine similarity with class-conditional Gaussian distributions. ML-AMCoS is designed to capture subtle nuances in feature covariance when intra-class variance is low, while suppressing covariance noise when the variance is high. To optimize the utilization of each layer, we calculate contribution weights based on performance against pseudo-OOD samples, which are generated by cut-mixing in-distribution (ID) training images with the four corners of images from different classes. Extensive experiments across diverse domains demonstrate that our method significantly improves robustness and detection accuracy, achieving an average AUROC of 87.13% and FPR@95 of 40.12% on near-OOD detection, outperforming the state-of-the-art methods by 7.25% and 7.08%, respectively.

1 Introduction

Despite deep learning's success in visual recognition, most classifiers assume a closed-world setting, leading to performance issues when exposed to out-of-distribution (OOD) inputs in open-world scenarios [1]. Such cases can lead to overconfident misclassifications, posing risks in safety-critical domains such as autonomous driving [2] and medical diagnostics [3]. OOD detection aims to

*Corresponding Author

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



(a) (b)

Figure 1: Near-OOD performance vs. layer depth (a) and feature variance (b). Optimal detection is domain-variant: Mahalanobis distance excels in medical domains (MIDOG, PhaKIR), while cosine similarity favors natural images (CIFAR-100, ImageNet-1k). The right plot suggests that Mahalanobis distance performs well when feature variance is low, as indicated by the log trace of the covariance matrix $\hat{\Sigma}$. However, it is worse than cosine similarity when variance increases. Our AMCoS leverages this correlation to adaptively bridge both metrics for robust cross-domain OOD detection.

identify whether an input originates from the same distribution as the training data, thus serving as a crucial safeguard for reliable AI deployment. In a nutshell, a typical OOD detector constructs a scoring function $\phi(x)$ that estimates how likely an input feature x belongs to the in-distribution (ID) using a threshold $\phi(x) \geq \tau$ to separate ID and OOD samples. Among existing designs, training-free methods are popular for their efficiency and model-agnostic nature. They construct $\phi(x)$ by exploiting statistical or geometric cues of pretrained features, typically gathered from the penultimate layer.

While training-free approaches achieve remarkable benchmarks on general-domain image classification tasks such as CIFAR [4] and ImageNet [5], their performance often degrades significantly in other domains, such as medical image classification. The key reason for this performance gap lies in the diversity of OOD examples. Some OOD samples are difficult to identify using high-level representations such as logits or deep features, yet they are more easily distinguished using representations from shallower layers. To illustrate this, we evaluate the near-OOD performance of Mahalanobis distance (Euclidean distance under class-conditional Gaussian distributions) and cosine similarity by using both penultimate features and features extracted from shallower layers across distinctive domains: CIFAR-100, ImageNet-1k, MIDOG [6] (pathology images of tumor tissue), and PhaKIR [7] (frames of surgery video). Following existing benchmarks [8, 9], we extract embeddings using their corresponding pretrained models from three depths: 1) shallow layer: the output after the initial convolution layer, 2) intermediate layer: the second residual block output, and 3) deep layer: the penultimate representation after global average pooling of the last residual block. The results are shown in Fig. 1.

We make two observations from Fig. 1-(a). First, using shallow features drops performance in natural image datasets (e.g., CIFAR-100 and ImageNet-1k) but yields substantial gains in medical image datasets (e.g., MIDOG and PhaKIR), while deep features show the opposite. This underscores that the effectiveness of OOD detection depends on domain-variant discriminability residing in different layers. Second, the Mahalanobis distance demonstrates inconsistent performance across domains, achieving an AUROC exceeding 90% in medical datasets while falling below 60% on natural images. A possible explanation is that the discriminant information in feature variance is more stable in medical domains as the average variance is relatively small as shown in Fig. 1-(b), possibly due to medical imaging acquisition usually following standard procedure, resulting in less data variance; whereas a high variance structure can become noise [10, 11], leading to overfitting. Under such circumstances, cosine similarity serves as a more robust alternative; given that classification models typically exhibit "Neural Collapse" [12], features are naturally trained to align with class weights, resulting in high dot products with their respective class centroids from each layer.

Motivated by these observations, we propose Multi-Layer Adaptive Mahalanobis-Cosine Similarity (ML-AMCoS), which integrates the strengths of Mahalanobis distance and cosine similarity through adaptive weighting based on the trace of the feature covariance at each layer. This mechanism

enables the detector to capture subtle structural nuances in feature distributions when the trace is low, while effectively suppressing covariance noise when the trace is high. As demonstrated in Fig. 1-(b), the *single-layer* version of our method, AMCoS, effectively bridges these two metrics to achieve superior robustness. Furthermore, to address the challenge of cross-domain layer selection, we implement a performance-driven weighting strategy. By generating pseudo-OOD samples through corner-cutmixing, which blends ID images with the corner regions of samples from other classes, we determine optimal contribution weights for each layer without requiring real OOD data.

In summary, our main contributions are as follows: 1) We propose ML-AMCoS, an OOD detection scoring function that adaptively combines cosine similarity and class-conditional Gaussian distributions. 2) We implement an automated strategy to optimize the utilization of representations across different layers, allowing the framework to adapt to domain-specific discriminative needs without requiring real OOD samples. 3) We show extensive evidence on a comprehensive set of OOD benchmarks that our approach achieves an average AUROC of 87.13% and FPR@95 of 40.12% on near-OOD detection, outperforming the state-of-the-art methods by 7.25% and 7.08%, respectively.

2 Related Work

Training-free OOD Detection. Training-free methods identify OOD samples using model outputs or features without backbone retraining. These fall into three groups: 1) Output-based methods leverage logits or probabilities, such as MSP [13], MaxLogit [14], energy scores [15], and perturbation-based ProGen [16]. 2) Single-layer feature methods utilize penultimate representations via Mahalanobis distance [17], KNN [18], neural collapse [19], or residual covariance [11], often refined by weight sparsification (DICE [20]) or residual integration (ViM [21]). 3) Multi-layer methods aggregate discriminability across network depths. While Gram [22], MBM [23], and S&I [24] use uniform summation, more advanced selection involves supervised real-OOD data (MDSEns [17]) or pseudo-OOD proxies like Jigsaw puzzles (FeatureNorm [25]). Despite these advances, existing methods rely on fixed metrics that fail to adapt to varying layer-wise or domain-specific feature distributions. We address this by introducing an adaptive scoring function and a performance-driven weighting strategy that optimizes multi-layer utilization without requiring real OOD samples or post-training.

Training-based OOD Detection. Training-based OOD methods aim to make OOD samples more distinguishable directly during training by modifying the learning process. Strategies include adding auxiliary network branches to estimate uncertainty [26, 27], applying data augmentation to improve robustness [28], calculating differences in activated regions between independent models [29], applying contrastive learning to obtain more discriminative embeddings [30], or adding energy regularization to explicitly enhance separation between ID and OOD samples [31]. The common assumption is that improved training objectives or architectures yield representations where ID and OOD distributions become more separable. However, training-based approaches typically introduce additional computational overhead and, in some cases, require access to auxiliary OOD data, which can limit their practicality in real-world deployments.

3 Preliminary

Given a pre-trained encoder, we can obtain a set of ID features $f_{id} \in \mathcal{F}_{id}$ and OOD features $f_{ood} \in \mathcal{F}_{ood}$ from the training data. Considering an anchor feature f_a, defined as the cluster center of ID features, an intuitive assumption is that $dist(f_{id}, f_a) \leq dist(f_{ood}, f_a)$, where $dist(\cdot, \cdot)$ denotes a distance metric. Compared to the Euclidean distance, which measures deviations from the mean without considering the underlying feature distribution, the Mahalanobis distance has been shown to be more effective for OOD detection [17, 32], as it incorporates the covariance structure and thus captures the geometry of the feature space. Formally, the Mahalanobis distance can be expressed as a matrix-induced distance:

$$dist(f, f_a) = \sqrt{(f - f_a)^\top M (f - f_a)}, \tag{1}$$

where M defines the geometry of the feature space. When M is the identity matrix, the metric reduces to the standard Euclidean distance.

Existing training-free methods primarily differ in how they define M. Lee *et al.* [17] defined M as the shared precision matrix Σ^{-1} of the ID class distributions, assuming a class-conditional Gaussian

structure. Building on this, ViM [21] defines $M = R^\top R$, where R represents the basis of the residual subspace, while NeCo [33] utilizes $M = P^\top P$ derived from the principal subspace P. However, a significant limitation of these approaches is their reliance on a static geometry M predominantly optimized for the penultimate layer. While Lee *et al.* [17] attempted to aggregate scores across multiple layers, their method requires real OOD samples and additional post-training procedures.

4 Methodology

As shown in Fig. 1, different metrics offer distinct advantages for OOD detection depending on the specific domain and layer depth. To develop a generalized framework, we propose Multi-Layer Adaptive Mahalanobis-Cosine Similarity (ML-AMCoS), which effectively uses the complementary strengths of the Mahalanobis distance and cosine similarity while accounting for layer-wise discriminative power. In this section, we first introduce Adaptive Mahalanobis-Cosine Similarity (AMCoS), a scoring function that dynamically adjusts the influence of feature covariance based on layer-specific statistics. Next, we describe our performance-driven layer selection strategy, which utilizes a corner-cutmixing technique to optimize layer weights without requiring real OOD data.

4.1 Adaptive Mahalanobis-Cosine Similarity (AMCoS)

For features extracted from any given layer l of a model with L layers, we aim to exploit the discriminative cues embedded within the feature covariance. We first compute the empirical class means and the tied covariance matrix from the training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$. For a specific layer l, the class mean $\hat{\mu}_c^l$ and the shared covariance $\hat{\Sigma}^l$ are defined as:

$$\hat{\mu}_c = \frac{1}{N_c} \sum_{i:y_i=c} f(\mathbf{x}_i), \quad \hat{\Sigma} = \frac{1}{N} \sum_c \sum_{i:y_i=c} (f(\mathbf{x}_i) - \hat{\mu}_c)(f(\mathbf{x}_i) - \hat{\mu}_c)^\top, \tag{2}$$

where N_c denotes the number of training samples for class c. For clarity, we omit the superscript l from these quantities and will reintroduce it later.

Traditional Mahalanobis-based methods rely on a static precision matrix, which can introduce significant noise in layers where features exhibit high variance. In such cases, incorporating the covariance matrix offers little additional discriminative value and may instead degrade performance [10, 11]. To address this, AMCoS calculates the similarity between a test sample $\mathbf{x}$ and the *nearest* class-conditional distribution using an adaptive scoring function for each layer:

$$\text{AMCoS}(x) = \max_c \frac{f(\mathbf{x})^T M \hat{\mu}_c}{\sqrt{f(\mathbf{x})^T M f(\mathbf{x})} \sqrt{\hat{\mu}_c^T M \hat{\mu}_c}}, \quad M = \lambda \hat{\Sigma}^{-1} + (1-\lambda) I. \tag{3}$$

The weight λ controls the contribution of the covariance matrix relative to the identity matrix. Since high-variance layers are prone to noise, we derive λ from the trace of the covariance square root:

$$\lambda = \exp\left(-\frac{\text{tr}(\sqrt{\hat{\Sigma}^l})}{\min\limits_{l' \in \{1,\dots,L\}} \text{tr}(\sqrt{\hat{\Sigma}^{l'}})}\right), \tag{4}$$

where $\min_{l' \in \{1,\dots,L\}} \text{tr}(\sqrt{\hat{\Sigma}^{l'}})$ serves as a normalization constant across layers. By formulation, $\lambda \in (0, e^{-1}]$, reaching its upper bound $\lambda \approx 0.368$ when the layer trace attains the global minimum across the network. While $\lambda \leq 0.368$, $\hat{\Sigma}^{-1}$ operates on normalized feature spaces where its elements scale inversely with feature variance. Since empirical intra-class variances are typically small (e.g., $\text{tr}(\hat{\Sigma}) \approx 10^{-3}$, yielding entries in $\hat{\Sigma}^{-1}$ on the order of 10^3), a weight of $\lambda \approx 0.368$ provides decisive numerical dominance to $\hat{\Sigma}^{-1}$ over the unit diagonal of I in low-variance layers. Conversely, in layers with high intra-class variance where the trace increases, $\lambda \to 0$, effectively suppressing noisy covariance structures and smoothly reverting M to the identity matrix I, which reduces the metric to robust cosine similarity.

4.2 Layer Selection via Corner-Cutmixing

To optimize the utilization of representations across different depths, we implement a performance-driven weighting strategy. Given a validation set of ID training samples, we generate pseudo-OOD

Table 1: Comparison of OOD detection methods on **cs-ID** across domains. Our method achieves the best overall performance. **Bold** indicates the best results, and underline denotes the second-best. CIF and IMN denote CIFAR and ImageNet, respectively.

Benmarks	OpenOOD				OpenMIBOOD			Average of both		
	CIF-10	CIF-100	IMN-200	IMN-1K	MIDOG	PhaKIR	OASIS-3	$\overline{\text{AUROC}}\uparrow$	$\overline{\text{AUPR}}\uparrow$	$\overline{\text{FPR@95}}\downarrow$
MSP [13]	75.51	78.18	74.43	74.00	53.37	32.04	36.95	60.64	54.91	84.12
ODIN [35]	68.85	78.81	72.95	75.30	56.38	34.28	53.51	62.87	57.24	83.39
MDSEns [17]	82.52	74.98	68.22	68.50	99.19	65.05	**100.00**	79.78	79.04	54.47
Gram [22]	**84.46**	**84.14**	75.27	73.43	73.38	20.31	99.79	72.97	71.37	62.21
KNN [18]	79.55	81.10	75.63	78.09	57.97	34.05	98.78	72.17	65.11	68.75
MBM [23]	80.09	77.08	68.83	70.98	98.90	**65.55**	**100.00**	80.20	77.26	68.75
FeatureNorm [25]	81.15	65.82	51.05	63.60	59.03	46.99	40.28	58.27	52.62	86.03
Dynamic [11]	77.29	79.18	75.12	77.34	52.26	50.60	99.91	73.10	72.05	70.12
S&I [24]	81.62	76.03	72.84	71.61	49.33	51.31	51.77	64.93	57.94	82.62
ProGen [16]	79.30	79.79	75.80	77.62	53.43	50.00	32.33	64.04	57.07	81.76
NCI [19]	76.86	80.71	76.11	77.34	57.78	52.71	45.33	66.69	60.02	77.20
ML-AMCoS	77.64	82.76	**76.13**	**78.22**	**99.22**	65.10	**100.00**	**82.72**	**79.99**	**52.45**

samples using a corner-cutmixing technique [34]. Specifically, each ID sample is mixed with four corner square patches (each of size $\alpha H \times \alpha W$, where α is a hyper-parameter) extracted from an ID image of a different class.

We evaluate the detection accuracy ACC^l of the AMCoS score for each layer l against these synthesized samples. To prioritize the most discriminative layers and suppress noise from sub-optimal ones, we retain only the top-k performing layers. Let $\mathcal{K}$ be the set of indices for the top-k layers based on ACC^l. We define the normalized weights w^l using a masked Softmax operation:

$$w^l = \frac{\exp(ACC^l)}{\sum_{l' \in \mathcal{K}} \exp(ACC^{l'})} \cdot \mathbb{1}(l \in \mathcal{K}), \tag{5}$$

where $\mathbb{1}(\cdot)$ is the indicator function that is 1 if $l \in \mathcal{K}$ and 0 otherwise. This ensures that only the most reliable layers contribute to the final OOD detection confidence score $\phi(\mathbf{x})$, which is calculated as:

$$\phi(\mathbf{x}) = \text{ML-AMCoS}(\mathbf{x}) = \sum_{l \in \mathcal{K}} w^l \text{AMCoS}^l(\mathbf{x}) \tag{6}$$

This selection mechanism allows the framework to adaptively prioritize the most relevant layers for a given domain without requiring real OOD data or manual layer tuning.

5 Experiments

5.1 Benchmarks

To evaluate our method, we conduct experiments on two open-source OOD detection benchmarks: OpenOOD [8] and OpenMIBOOD [9].

OpenOOD contains 4 natural image datasets, CIFAR-10, CIFAR-100, ImageNet-200, and ImageNet-1K. Each dataset includes 15 common corruptions to simulate covariate shifts (cs-ID). Near-OOD samples share visual structure with ID classes (e.g., CIFAR-100 for CIFAR-10, SSB-Hard for ImageNet-200), whereas far-OOD datasets such as MNIST, SVHN, and iNaturalist introduce larger domain and semantic gaps. Following the benchmark protocol, we use ResNet-18 for CIFAR-10/100 and ImageNet-200, and ResNet-50 for ImageNet-1K.

OpenMIBOOD evaluates cross-domain generalization in medical OOD detection across pathology images (MIDOG), surgical videos (PhaKIR), and brain MRI (OASIS-3). Each dataset defines cs-ID, near-OOD, and far-OOD based on domain variations, such as scanner or staining changes, inter-dataset shifts, or cross-modality differences. We follow the official setup using ResNet-50 for MIDOG, ResNet-18 for PhaKIR, and R(2+1)D for OASIS-3.

A more detailed summary of the experimental setup, covering all ID, cs-ID, near-OOD, and far-OOD datasets, as well as the model backbones used in our study, is presented in Appx. A.1.

Table 2: Comparison of OOD detection methods on **near-OOD** across domains using AUROC unless otherwise specified. MS-AMCoS achieves the best overall performance. **Bold** and underline indicate the 1st and 2nd places in each column. CIF and IMN denote CIFAR and ImageNet, respectively.

Benchmarks	**OpenOOD**				**OpenMIBOOD**			**Averages for both**		
	CIF-10	CIF-100	IMN-200	IMN-1K	MIDOG	PhaKIR	OASIS-3	$\overline{\text{AUROC}}\uparrow$	$\overline{\text{AUPR}}\uparrow$	$\overline{\text{FPR@95}}\downarrow$
MSP [13]	87.68	80.42	83.30	76.02	55.90	50.16	53.50	69.57	59.48	70.02
ODIN [35]	80.25	79.80	80.32	74.75	61.37	41.78	59.53	68.26	60.68	77.50
MDSEns [17]	88.16	61.92	64.85	55.82	<u>91.84</u>	**97.11**	99.46	<u>79.88</u>	<u>75.37</u>	<u>47.20</u>
Gram [22]	87.25	76.03	73.47	64.28	75.72	57.63	99.58	76.28	70.00	64.12
KNN [18]	**90.70**	80.25	81.75	71.10	61.63	55.44	97.66	76.93	69.57	59.28
MBM [23]	76.38	51.93	62.18	55.94	91.47	<u>94.84</u>	<u>99.97</u>	76.10	71.75	53.19
FeatureNorm [25]	73.22	48.69	58.83	68.54	51.62	34.01	53.19	55.44	47.07	88.82
Dynamic [11]	88.65	76.97	83.38	74.38	52.59	61.60	99.59	76.74	74.57	61.27
S&I [24]	90.53	72.78	73.12	64.13	54.69	72.16	68.67	70.87	67.19	71.15
ProGen [16]	<u>90.61</u>	**81.33**	<u>83.72</u>	<u>76.81</u>	55.74	59.02	46.46	70.53	66.50	68.31
NCI [19]	88.75	81.04	83.55	**78.69**	60.81	69.04	49.61	73.07	69.18	67.38
ML-AMCoS	89.10	<u>81.06</u>	**83.82**	72.52	**91.90**	91.49	**99.99**	**87.13**	**82.14**	**40.12**

5.2 Metrics and Compared Methods

We follow the evaluation protocol established in two benchmarks, where OOD samples are treated as the positive class. The primary metric is the area under the receiver operating characteristic curve (AUROC), which measures the overall separability between ID and OOD samples across all thresholds. We also report the false positive rate at 95% true positive rate (FPR@95), which quantifies the proportion of ID samples misclassified as OOD when 95% of OOD samples are correctly identified. In addition, we include the area under the precision–recall curve (AUPR) as a complementary metric, following the harmonic mean formulation proposed in OpenMIBOOD:

$$\text{AUPR} = \frac{2 \cdot (\text{AUPR}_{\text{ID}} \cdot \text{AUPR}_{\text{OOD}})}{\text{AUPR}_{\text{ID}} + \text{AUPR}_{\text{OOD}}},$$

where AUPR_{ID} is computed by treating ID samples as the positive class, reflecting the model's ability to correctly identify ID data, while AUPR_{OOD} treats OOD samples as positive, evaluating the model's capacity to detect unseen or unexpected inputs.

We compare our method with a diverse set of training-free state-of-the-art OOD detection approaches, including three output-based method methods (MSP, ODIN, ProGen), three feature-based methods (KNN, Dynamic, NCI), and five multi-layer methods (MDSEns, MBM, FeatureNorm, Gram, S&I). For all multi-layer methods, we follow the implementation of [17] and extract features from every end of residual block of ResNet.

5.3 Experimental Results

Tables 1 and 2 compare our method with previous training-free OOD detection approaches on cs-ID and near-OOD tasks, which are more challenging than far-OOD scenarios, which are included in the Appx. A.3 for completeness. While baseline hyperparameters are tuned using held-out near-OOD validation sets, our method consistently utilizes the top-2 layers and an $\alpha = 0.3$ across all datasets. Our approach ranks first overall in AUROC, AUPR, and FPR@95, demonstrating strong robustness across diverse domains. In particular, it achieves outstanding performance on both cs-ID and near-OOD detection, with average AUROC scores of 82.72% and 87.13%, outperforming the second-best methods MBM and MDSEns by 2.52% and 7.25%, respectively. Notably, our method also achieves a substantial improvement in FPR@95, gaining 2.02% and 7.08% over MDSEns. This is particularly significant as MDSEns requires real near-OOD data for layer selection, whereas our approach achieves superior false-positive mitigation through a fully training-free, OOD-free strategy.

These results also highlight a critical discrepancy in existing literature. While methods like MDSEns and MBM attain high performance on the medical-centric OpenMIBOOD benchmark, they perform substantially worse on the natural images of OpenOOD. Conversely, methods optimized for OpenOOD often struggle with the domain-specific nuances of OpenMIBOOD. This suggests that existing approaches are frequently over-fitted to specific feature distributions. In contrast, ML-AMCoS

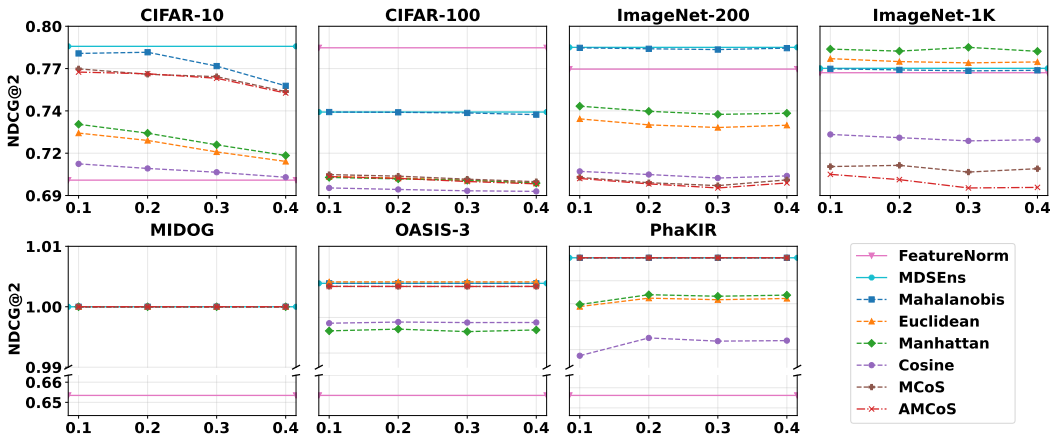


Figure 2: Layer weighting quality analysis using NDCG@2 across domains and metrics on **near-OOD** detection task. Our corner-cutmixing strategy ($\alpha \in [0.1, 0.4]$) is evaluated against MDSEns, which serves as a supervised upper bound for Mahalanobis-based selection, and FeatureNorm, which utilizes Jigsaw-based pseudo-OOD samples. Higher values indicate superior alignment between predicted layer weights w^l and ground-truth AUROC performance corresponding to each layer. Our strategy demonstrates significant robustness across diverse domains; notably, the Mahalanobis distance under our selection framework nearly reaches the performance of the supervised upper bound without requiring real OOD data.

demonstrates strong cross-domain generalizability by adaptively leveraging feature covariance and optimizing layer-wise contributions.

However, certain edge cases remain challenging. All evaluated methods exhibit relatively low performance on the PhaKIR cs-ID subset, primarily due to the temporal nature of smoke occurrences. When smoke is faint, it remains nearly imperceptible in static imagery and only becomes discernible through temporal comparisons with preceding frames (as detailed in the Appx. A.4). In single frames, these occurrences are often too subtle to significantly alter feature activations. Consequently, AUROC scores near 0.5 reflect an inherent lack of OOD cues in static data rather than a failure of the detection methods themselves. If the downstream task shifted toward organ classification, smoke would likely have a more pronounced impact by obscuring critical diagnostic features, thereby enhancing the discriminability of OOD samples [9].

5.4 Alignment between Layer Weighting w^l and near-OOD Performance

In this section, we analyze the alignment between our pseudo-OOD weighting strategy and the true near-OOD performance of various distance metrics; results for cs-ID and far-OOD tasks are provided in Appx. A.5. We evaluate the layer-wise ranking quality of Euclidean, Mahalanobis, Manhattan distances, and cosine similarity, alongside *singel-version* of our method (AMCoS) and its non-adaptive variant (MCoS), which excludes the adaptive weighting parameter λ. We compare our strategy against the selection methods proposed in MDSEns [17] and FeatureNorm [25] across varying values of hyperparameter α.

Since OOD discriminability varies drastically across network depths, a global metric such as Pearson correlation is insufficient for evaluating a strategy's ability to prioritize the most discriminative layers. We thus utilize Normalized Discounted Cumulative Gain (NDCG@k), commonly used in information retrieval, to evaluate ranking quality. To specifically account for the magnitude of our derived weights w^l, we utilize a weighted formulation of DCG:

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}, \quad \text{where DCG@}k = \sum_{i=1}^{k} \frac{w^i(2^{rel^i} - 1)}{\log_2(i+1)}. \tag{7}$$

Here, rel^i represents the ground-truth AUROC of the i-th ranked layer, and w^i is normalized across layers using Softmax function. For benchmarking, we use the layer selection criteria proposed by

MDSEns and FeatureNorm as their respective w^i, paired with their native distance metrics for rel^i (Mahalanobis distance for MDSEns and the norm of the feature map for FeatureNorm). In our experiments, we report results with $k = 2$, maintaining consistency with the top-2 setting employed in our main results.

As shown in Fig. 2, our weighting strategy generalizes well across diverse metrics and domains. Notably, when utilizing Mahalanobis distance, our NDCG@2 scores are nearly indistinguishable from those of MDSEns, an upper bound for Mahalanobis-based layer selection, across most datasets. This is significant, as our strategy achieves the same ranking precision without requiring real OOD validation samples. Furthermore, while FeatureNorm achieves competitive ranking on CIFAR-100, its performance exhibits catastrophic instability in other domains, particularly medical benchmarks such as MIDOG, OASIS-3, and PhaKIR. In contrast, our strategy maintains a near-perfect NDCG across all medical datasets. This stability across $\alpha \in [0.1, 0.4]$ indicates that corner-cutmixing consistently identifies the most discriminative feature scales, regardless of whether the distance metric is geometric or statistical. However, high alignment does not necessarily result in superior OOD detection performance. As illustrated in Fig. 1, the absolute performance of the Mahalanobis metric is lower than that of AMCoS on natural images. Thus, even though our layer weighting achieves higher alignment with the Mahalanobis distance compared to AMCoS on natural domains, the final performance will not exceed that of the full ML-AMCoS framework. Ultimately, this confirms that ML-AMCoS provides a reliable mechanism for layer prioritization that outperforms existing OOD-free strategies (FeatureNorm) while matching the precision of supervised selection (MDSEns). Finally, we report results for a center-cutmixing variant in Appx. A.5, which utilizes center-block perturbations rather than corner-based shifts.

5.5 Ablation Study

To verify the effectiveness of components, we conduct a comprehensive ablation study across seven datasets for both cs-ID and near-OOD tasks (the results of far-OOD is included in the Appx. A.6). We split the ResNet-50 backbone into five distinct layers: Layer 1 (initial convolution), Layers 2–4 (outputs of the first three residual blocks), and Layer 5 (the penultimate layer).

Analysis of Layer Selection ($\mathcal{K}$). As detailed in Tab. 3 and 4, while performance remains relatively stable across most natural image datasets when increasing the number of aggregated layers, we observe a performance collapse in the medical domain, particularly on PhaKIR. Specifically, AUROC scores drop by over 20% and 30% in cs-ID and near-OOD tasks, respectively, when expanding the selection beyond the top-2 layers. This phenomenon suggests that in specialized domains, discriminative OOD cues are highly localized within specific semantic scales. Including deeper or shallower layers indiscriminately introduces feature dilution, where non-informative noise from sub-optimal layers overwhelms the discriminative signal. Our top-k selection strategy acts as a critical information bottleneck, preserving only the most domain-relevant representations.

The Effect of Layer Weighting (w^l). Replacing our performance-driven weights with a uniform average (w/o w^l) leads to a consistent degradation in near-OOD detection across both benchmarks and cs-ID tasks on OpenMIBOOD, with the PhaKIR benchmark exhibiting the most severe sensitivity. As shown in Tab. 3 and 4, the AUROC for PhaKIR collapses by over 24% and 32% in cs-ID and near-OOD tasks, respectively, when weights are removed. This drastic decline suggests that for domains where OOD cues are extremely subtle, such as faint smoke in surgical scenes (see Appx. A.4), the discriminative signal is highly concentrated in specific layers. A simple uniform average effectively washes out these sparse cues by assigning equal importance to all layers, including non-informative ones. By utilizing corner-cutmixing to synthesize pseudo-OOD samples, ML-AMCoS successfully identifies and prioritizes the layers where these subtle manifold shifts are detectable, providing a robust weighting scheme that avoids the need for real-world OOD samples.

Importance of Adaptive Covariance (λ). The removal of the adaptive mechanism (w/o λ) results in a performance decline across nearly all benchmarks, with ImageNet-1K exhibiting a loss of over 3%. This result provides empirical evidence for our core hypothesis: a static precision matrix is insufficient for multi-layer aggregation. Specifically, in layers where the feature covariance exhibits a high trace, the precision matrix Σ^{-1} becomes numerically unstable or over-fits to ID-specific noise, which can misalign OOD features and degrade detection stability. Our adaptive λ addresses this by dynamically scaling the influence of the covariance based on its internal variance. By shifting toward the robustness of cosine similarity when variance noise is high, and leveraging fine-grained

Table 3: Ablation analysis of ML-AMCoS configurations on **cs-ID** detection task. The default setting (top-2) is marked with an asterisk (*). Performance variations relative to the default setting are indicated in the format AUROC$_{(\Delta\text{AUROC})}$. Relative improvements are highlighted in red, while performance decreases are indicated in teal.

	CIF-10	CIF-100	IMN-200	IMN-1K	MIDOG	PhaKIR	OASIS-3
ML-AMCoS (top-1)	77.92$_{(+0.28)}$	79.86$_{(-2.90)}$	75.03$_{(-1.10)}$	77.90$_{(-0.32)}$	99.21$_{(-0.01)}$	63.65$_{(-1.45)}$	100.00
ML-AMCoS (top-2)*	77.64	82.76	76.13	78.22	99.22	65.10	100.00
ML-AMCoS (top-3)	78.68$_{(+1.04)}$	83.32$_{(+0.56)}$	76.70$_{(+0.57)}$	78.31$_{(+0.09)}$	99.27$_{(+0.05)}$	40.87$_{(-24.23)}$	100.00
ML-AMCoS (top-4)	79.03$_{(+1.39)}$	83.47$_{(+0.71)}$	76.74$_{(+0.61)}$	78.33$_{(+0.11)}$	94.57$_{(-4.65)}$	40.89$_{(-24.21)}$	96.12$_{(-3.88)}$
ML-AMCoS (top-5)	79.08$_{(+1.44)}$	75.80$_{(-6.96)}$	76.92$_{(+0.79)}$	78.36$_{(+0.14)}$	95.56$_{(-3.66)}$	41.18$_{(-23.92)}$	96.25$_{(-3.75)}$
ML-AMCoS (w/o w^l)	79.51$_{(+1.87)}$	84.00$_{(+1.24)}$	77.02$_{(+0.89)}$	78.14$_{(-0.08)}$	95.73$_{(-3.49)}$	40.78$_{(-24.21)}$	95.73$_{(-4.27)}$
ML-AMCoS (w/o λ)	77.61$_{(-0.03)}$	81.15$_{(-1.61)}$	75.18$_{(-0.95)}$	75.15$_{(-3.07)}$	96.94$_{(-2.28)}$	65.00$_{(-0.10)}$	100.00

Table 4: Ablation analysis of ML-AMCoS configurations on **near-OOD** detection task. The default setting (top-2) is marked with an asterisk (*). Performance variations relative to the default setting are indicated in the format AUROC$_{(\Delta\text{AUROC})}$. Relative improvements are highlighted in red, while performance decreases are indicated in teal.

	CIF-10	CIF-100	IMN-200	IMN-1K	MIDOG	PhaKIR	OASIS-3
ML-AMCoS (top-1)	89.27$_{(+0.17)}$	79.58$_{(-1.48)}$	83.12$_{(-0.70)}$	73.72$_{(+1.20)}$	92.27$_{(+0.37)}$	89.91$_{(-1.58)}$	99.99
ML-AMCoS (top-2)*	89.10	81.06	83.82	72.52	91.90	91.49	99.99
ML-AMCoS (top-3)	89.15$_{(+0.05)}$	80.67$_{(-0.39)}$	83.20$_{(-0.62)}$	72.45$_{(-0.07)}$	91.57$_{(-0.33)}$	58.28$_{(-33.21)}$	99.99
ML-AMCoS (top-4)	89.21$_{(+0.11)}$	80.50$_{(-0.56)}$	83.17$_{(-0.65)}$	72.28$_{(-0.24)}$	93.40$_{(+1.50)}$	58.85$_{(-32.64)}$	94.79$_{(-5.20)}$
ML-AMCoS (top-5)	89.25$_{(+0.15)}$	80.48$_{(-0.58)}$	83.12$_{(-0.70)}$	72.24$_{(-0.28)}$	94.29$_{(+2.39)}$	59.56$_{(-31.93)}$	94.98$_{(-5.01)}$
ML-AMCoS (w/o w^l)	88.80$_{(-0.30)}$	79.69$_{(-1.37)}$	81.82$_{(-2.00)}$	71.00$_{(-1.52)}$	94.43$_{(-2.53)}$	58.93$_{(-32.56)}$	94.43$_{(-5.56)}$
ML-AMCoS (w/o λ)	88.69$_{(-0.41)}$	78.94$_{(-2.12)}$	82.58$_{(-1.24)}$	69.29$_{(-3.23)}$	96.51$_{(+4.61)}$	91.53$_{(+0.04)}$	99.99

Mahalanobis cues only when they are reliably estimated, ML-AMCoS maintains an optimal balance between discriminative power and geometric stability.

5.6 Evaluation Across Different Backbones

To evaluate the architectural robustness of ML-AMCoS, we compare it against four competitive training-free baselines using the ImageNet-1K benchmark across three distinct backbones: ResNet-50, ViT-B-16 [36], and Swin-T [37]. To ensure a fair comparison, we utilize pretrained checkpoints from the *torchvision* library, maintaining consistency with the OpenOOD evaluation protocol. For multi-layer extraction within transformer architectures, we extract features after each hierarchical stage in Swin-T, and from layers 1, 4, 7, 10, and the penultimate representation in ViT-B-16.

As shown in Tab. 5, ML-AMCoS demonstrates significant cross-architecture generalizability. Notably, our method achieves state-of-the-art performance across all three OOD scenarios (cs-ID, near-OOD, and far-OOD) when utilizing the ViT-B-16 backbone. While certain baselines show localized strengths, such as NCI on ResNet-50 near-OOD, ML-AMCoS remains the most consistently high-performing method across both convolutional and transformer-based models. These results suggest that our adaptive weighting strategy successfully extracts discriminative cues from diverse feature hierarchies, proving that the efficacy of multi-layer covariance adaptation is not confined to traditional CNN structures but extends to the global and hierarchical self-attention mechanisms of Transformers.

6 Conclusion

In this paper, we presented ML-AMCoS, a fully training-free framework designed to enhance OOD detection across diverse domains without requiring real OOD validation data. Our approach addresses two critical limitations of existing methods: the instability of feature covariance across network depths and the difficulty of selecting discriminative layers for specialized datasets. By introducing an adaptive Mahalanobis-Cosine similarity, ML-AMCoS dynamically balances statistical precision with geometric robustness based on the local feature distribution. Furthermore, our pseudo-OOD weighting strategy utilizing corner-cutmixing provides a reliable mechanism for identifying

Table 5: AUROC performance comparison of training-free OOD detection methods on ImageNet-1K using different backbones (ResNet-50, ViT-B-16, and Swin-T). Best results are highlighted in **bold**. Our method performs consistently well across architectures and achieves the best results on ViT-B-16, demonstrating its strong generalization beyond CNNs.

	ResNet-50			**ViT-B-16**			**Swin-T**		
	cs-ID	near-OOD	far-OOD	cs-ID	near-OOD	far-OOD	cs-ID	near-OOD	far-OOD
KNN [18]	78.09	71.10	90.18	73.93	74.11	90.81	74.86	71.62	89.37
MSP [13]	74.00	76.02	85.23	71.20	73.53	86.04	72.98	**76.75**	86.3
ProGen [16]	77.62	76.81	89.32	75.07	76.35	91.22	**77.49**	76.07	90.53
NCI [19]	77.34	**78.69**	**95.53**	73.49	73.72	88.88	75.18	75.93	90.11
ML-AMCoS	**78.22**	72.52	94.34	**75.33**	**77.82**	**92.39**	75.36	74.46	**92.12**

informative layers, achieving ranking precision comparable to supervised methods. Extensive evaluations on natural and medical imaging benchmarks demonstrate that ML-AMCoS consistently outperforms state-of-the-art training-free detectors. While our framework provides robust layer prioritization, a rigorous mathematical proof establishing this theoretical link is currently lacking and remains an open challenge for future work. Additionally, the framework's overall detection capability remains sensitive to the quality of the selected representations, future efforts could also focus on enhancing tolerance to potential layer mis-selection.

References

[1] Nick Drummond and Rob Shearer. The open world assumption. In *eSI workshop: the closed world of databases meets the open world of the semantic web*, volume 15, page 1, 2006.

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

[3] Wei Fu, Yufei Chen, Wei Liu, Xiaodong Yue, and Chao Ma. 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, 2023.

[4] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. `https://www.cs.toronto.edu/~kriz/cifar.html`, 2009. Accessed: 2025-09-09.

[5] 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.

[6] Marc Aubreville, Frauke Wilm, Nikolas Stathonikos, Katharina Breininger, Taryn A Donovan, Samir Jabari, Mitko Veta, Jonathan Ganz, Jonas Ammeling, Paul J van Diest, et al. A comprehensive multi-domain dataset for mitotic figure detection. *Scientific data*, 10(1):484, 2023.

[7] Tobias Rueckert, Raphael Maerkl, David Rauber, Leonard Klausmann, Max Gutbrod, Daniel Rueckert, Hubertus Feußner, Dirk Wilhelm, and Christoph Palm. Video dataset for surgical phase, keypoint, and instrument recognition in laparoscopic surgery (phakir). 2025.

[8] 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. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*.

[9] Max Gutbrod, David Rauber, Danilo Weber Nunes, and Christoph Palm. Openmibood: Open medical imaging benchmarks for out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25874–25886, 2025.

[10] Henryk Maciejewski, Tomasz Walkowiak, and Kamil Szyc. Out-of-distribution detection in high-dimensional data using mahalanobis distance-critical analysis. In *International Conference on Computational Science*, pages 262–275. Springer, 2022.

[11] Kaiyu Guo, Zijian Wang, Tan Pan, Brian C. Lovell, and Mahsa Baktashmotlagh. Improving out-of-distribution detection via dynamic covariance calibration. In *Forty-second International Conference on Machine Learning*, 2025.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] Wenxi Chen, Raymond A Yeh, Shaoshuai Mou, and Yan Gu. Leveraging perturbation robustness to enhance out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4724–4733, 2025.

[17] 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.

[18] 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.

[19] 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.

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

[21] 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.

[22] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with gram matrices. In *International conference on machine learning*, pages 8491–8501. PMLR, 2020.

[23] Harry Anthony and Konstantinos Kamnitsas. On the use of mahalanobis distance for out-of-distribution detection with neural networks for medical imaging. In *International workshop on uncertainty for safe utilization of machine learning in medical imaging*, pages 136–146. Springer, 2023.

[24] Jiayu Zhang, Xinyi Wang, Zhibo Jin, Zhiyu Zhu, Jianlong Zhou, Fang Chen, and Huaming Chen. Splitting & integrating: Out-of-distribution detection via adversarial gradient attribution. In *Forty-second International Conference on Machine Learning*, 2025.

[25] Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, and Kyoobin Lee. Block selection method for using feature norm in out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15701–15711, 2023.

[26] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018.

[27] Qi Chen and Hu Ding. Dual energy-based model with open-world uncertainty estimation for out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25728–25737, 2025.

[28] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*.

[29] Alessandro Canevaro, Julian Schmidt, Mohammad Sajad Marvi, Hang Yu, Georg Martius, and Julian Jordan. Advancing out-of-distribution detection via local neuroplasticity. In *The Thirteenth International Conference on Learning Representations*.

[30] Shawn Li, Huixian Gong, Hao Dong, Tiankai Yang, Zhengzhong Tu, and Yue Zhao. Dpu: Dynamic prototype updating for multimodal out-of-distribution detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10193–10202, 2025.

[31] Yiting Chen, Qitian Wu, and Junchi Yan. Regularizing energy among training samples for out-of-distribution generalization. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

[33] Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection. *CoRR*, 2023.

[34] Fen Fang, Nhat M Hoang, Qianli Xu, and Joo-Hwee Lim. Data augmentation using corner cutmix and an auxiliary self-supervised loss. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 830–834. IEEE, 2023.

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

[36] 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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[37] 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.

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

[39] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

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

[41] 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.

[42] 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 7. Granada, 2011.

[43] 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.

[44] 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.

[45] 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.

[46] 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*.

[47] 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.

[48] 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.

[49] João Gustavo Atkinson Amorim, Luiz Antonio Buschetto Macarini, André Victória Matias, Allan Cerentini, Fabiana Botelho De Miranda Onofre, Alexandre Sherlley Casimiro Onofre, and Aldo Von Wangenheim. A novel approach on segmentation of agnor-stained cytology images using deep learning. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 552–557. IEEE, 2020.

[50] Joao Gustavo Atkinson Amorim, Luiz Antonio Buschetto Macarini, Andre Victoria Matias, Allan Cerentini, Fabiana Botelho De Miranda Onofre, Alexandre Sherlley Casimiro Onofre, and Aldo Von Wangenheim. A novel approach on segmentation of agnor-stained cytology images using deep learning. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*. IEEE, July 2020.

[51] Amartya Ranjan Saikia, Kangkana Bora, Lipi B Mahanta, and Anup Kumar Das. Comparative assessment of cnn architectures for classification of breast fnac images. *Tissue and Cell*, 57:8–14, 2019.

[52] Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

[53] Sebastian Bodenstedt, Max Allan, Anthony Agustinos, Xiaofei Du, Luis Garcia-Peraza-Herrera, Hannes Kenngott, Thomas Kurmann, Beat Müller-Stich, Sebastien Ourselin, Daniil Pakhomov, et al. Comparative evaluation of instrument segmentation and tracking methods in minimally invasive surgery. *arXiv preprint arXiv:1805.02475*, 2018.

[54] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*, 2020.

[55] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International conference on multimedia modeling*, pages 451–462. Springer, 2019.

[56] Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Soumali Roychowdhury, Xiaowei Hu, Gabija Maršalkaitė, Odysseas Zisimopoulos, Muneer Ahmad Dedmari, Fenqiang Zhao, Jonas Prellberg, et al. Cataracts: Challenge on automatic tool annotation for cataract surgery. *Medical image analysis*, 52:24–41, 2019.

[57] Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medrxiv*, pages 2019–12, 2019.

[58] Sook-Lei Liew, Bethany P Lo, Miranda R Donnelly, Artemis Zavaliangos-Petropulu, Jessica N Jeong, Giuseppe Barisano, Alexandre Hutton, Julia P Simon, Julia M Juliano, Anisha Suri, et al. A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. *Scientific data*, 9(1):320, 2022.

[59] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.

[60] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[61] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[62] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[63] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*, 2019.

[64] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical image analysis*, 69:101950, 2021.

[65] Ali Emre Kavur, M Alper Selver, Oğuz Dicle, Mustafa Barış, and N Sinem Gezer. Chaos-combined (ct-mr) healthy abdominal organ segmentation challenge data. *(No Title)*, 2019.

[66] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[67] Catalina Tobon-Gomez, Arjan J Geers, Jochen Peters, Jürgen Weese, Karen Pinto, Rashed Karim, Mohammed Ammar, Abdelaziz Daoudi, Jan Margeta, Zulma Sandoval, et al. Benchmark for algorithms segmenting the left atrium from 3d ct and mri datasets. *IEEE transactions on medical imaging*, 34(7):1460–1473, 2015.

A Technical Appendices and Supplementary Material

A.1 Experimental Setup

Table 6: Summary of datasets used in OOD detection benchmarks. Each benchmark includes ID, cs-ID, near-OOD, and far-OOD datasets following the official OpenOOD and OpenMIBOOD protocols.

Benchmark	ID Dataset	cs-ID	near-OOD	far-OOD	Backbone
OpenOOD [8]	CIFAR-10 [38]	15 corruptions [39]	CIFAR-100, Tiny ImageNet [40]	MNIST [41], SVHN [42], Textures [43], Places365 [44]	ResNet-18 [45]
	CIFAR-100 [4]		CIFAR-10, Tiny ImageNet	MNIST, SVHN, Textures, Places365	ResNet-18
	ImageNet-200 [5]		SSB-Hard [46], NINCO [47]	iNaturalist [48], Textures, OpenImage-O [21]	ResNet-18
	ImageNet-1K [5]		SSB-Hard, NINCO	iNaturalist, Textures, OpenImage-O	ResNet-50 [45]
OpenMIBOOD [9]	MIDOG [6]	Scanner/Staining shifts	Other domain shifts in MIDOG	CCAgT [49, 50], FNAC 2019 [51]	ResNet-50
	PhaKIR [7]	Smoke levels	Cholec80 [52], EndoSeg15 [53], EndoSeg18 [54]	Kvasir-SEG [55], CATARACTS [56]	ResNet-18
	OASIS-3 [57]	T2-weighted, Scanner shift	ATLAS [58], BraTS [59, 60, 61], OASIS-CT [57]	MSD-Heart [62, 63], CHAOS [64, 65]	R(2+1)D [66]

To validate the effectiveness of our method, we conduct experiments on two open-sourced OOD detection benchmarks: OpenOOD [8] and OpenMIBOOD [9]. A comprehensive overview of these benchmarks is presented in Tab. 6.

OpenOOD consists four ID datasets: CIFAR-10 [38], CIFAR-100 [4], ImageNet-200, and ImageNet-1K [5].

CIFAR-10 leverages CIFAR-100 and Tiny ImageNet [40] as near-OOD data, while MNIST [41], SVHN [42], Textures [43], and Places365 [44] are used as far-OOD data.

CIFAR-100 uses CIFAR-10 and Tiny ImageNet as near-OOD and the same far-OOD datasets as CIFAR-10.

ImageNet-200, a 200-class subset of ImageNet-1K, uses SSB-hard [46] and NINCO [47] as near-OOD sets, and iNaturalist [48], Textures, and OpenImage-O [21] as far-OOD datasets.

ImageNet-1K shares OOD sets with ImageNet-200.

In all cases, 15 types of image corruptions are applied to generate cs-ID samples following [39], and a small subset of the ID test set is reserved for validation. Following the benchmark protocol, we use ResNet-18 for CIFAR-10/100 and ImageNet-200, and ResNet-50 for ImageNet-1K.

OpenMIBOOD considers three datasets as ID sets: MIDOG [6], PhaKIR [7], and OASIS-3 [57].

MIDOG uses histopathology images for mitosis detection, where domain shifts caused by different scanners and staining protocols form cs-ID samples, other domains were classified as near-OOD, and cytology datasets (CCAgT [49, 50] and FNAC 2019 [51]) serve as far-OOD.

PhaKIR consists of laparoscopic surgery videos, with smoke-free frames as ID data, varying smoke levels as cs-ID, other surgical datasets (Cholec80 [52], EndoSeg15 [53], EndoSeg18[54]) as near-OOD, and visually distinct datasets (Kvasir-SEG [55] and CATARACTS [56]) as far-OOD.

OASIS-3 evaluates brain MRI classification, where T1-weighted scans constitute the ID set, T2-weighted and different-scanner data form cs-ID, and cross-disease or cross-modality datasets (ATLAS [58], BraTS [59, 60, 61], CT scans from OASIS-3) serve as near-OOD, while unrelated anatomical datasets (MSD-Heart [62, 63, 67] and CHAOS [64, 65]) are used as far-OOD.

All classifiers follow the benchmark protocol, using ResNet-50 for MIDOG, ResNet-18 for PhaKIR, and R(2+1)D [66] for OASIS-3. Regarding hardware, OASIS-3 experiments were conducted on an NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition (94G), while all other experiments were performed using NVIDIA GeForce RTX 4090 (24G) GPUs.

A.2 OOD Detection Performance Cross Layers

In this section, we extend the analysis introduced in Fig. 1 to provide a comprehensive view of layer-wise OOD discriminability across all seven benchmarks and three detection tasks: cs-ID (Fig. 3), near-OOD (Fig. 4), and far-OOD (Fig. 5).

Consistent with our preliminary findings, we observe a distinct divergence in optimal layer depth between natural and medical image domains. For natural image datasets (e.g., CIFAR-100 and ImageNet), the penultimate layers consistently yield superior performance. This suggests that

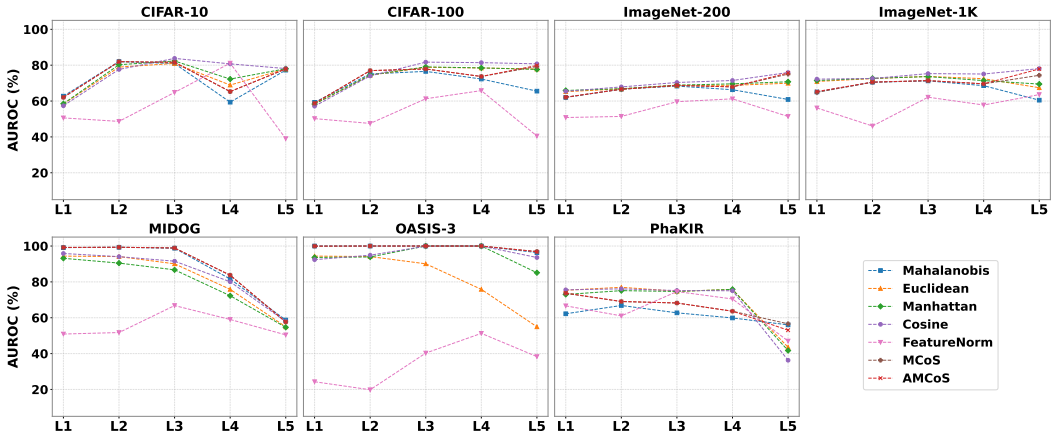


Figure 3: Layer-wise **cs-ID** detection performance. Shallow representations (Layer 1) achieve superior performance in medical domains (MIDOG, OASIS-3, and PhaKIR) but degrade in natural domains (CIFAR-10/100 and ImageNet-200/1k). Conversely, deep representations (Layer 5) exhibit the opposite trend, indicating that optimal OOD discriminability is highly layer-dependent.

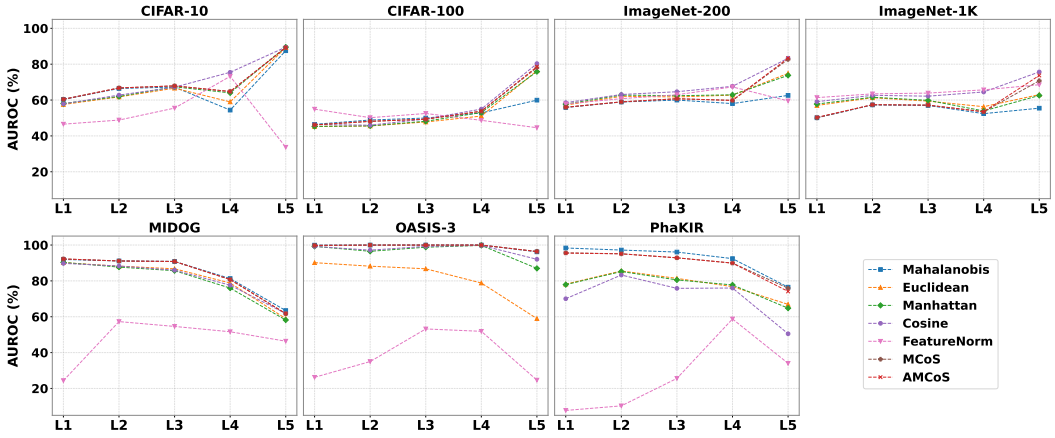


Figure 4: Layer-wise **near-OOD** detection performance. Shallow representations (Layer 1) achieve superior performance in medical domains (MIDOG, OASIS-3, and PhaKIR) but degrade in natural domains (CIFAR-10/100 and ImageNet-200/1k). Conversely, deep representations (Layer 5) exhibit the opposite trend, indicating that optimal OOD discriminability is highly layer-dependent.

semantically rich, high-level features are necessary to distinguish between natural object classes. Conversely, for medical benchmarks such as MIDOG and PhaKIR, shallow layers often exhibit significantly higher discriminability. This behavior indicates that OOD cues in specialized medical imaging are frequently captured by low-level textural or structural perturbations rather than high-level semantic shifts. Furthermore, on natural images where the Mahalanobis distance performs poorly, cosine similarity serves as a more robust alternative for OOD detection. Its superior performance over Euclidean and Manhattan distances in most instances suggests it is better suited to the feature geometry of classification tasks. In such settings, features are naturally trained to align with class weights, resulting in high dot products with their respective class centroids at each layer. Finally, the results highlight the limited generalizability of single-strategy baselines. For instance, FeatureNorm [25] demonstrates competitive results on natural image benchmarks but fails significantly on medical datasets. This lack of robustness underscores the necessity of a domain-agnostic approach like ML-AMCoS, which adaptively weights contributions across the entire feature hierarchy based on local statistical reliability rather than making static assumptions about optimal layer depth.

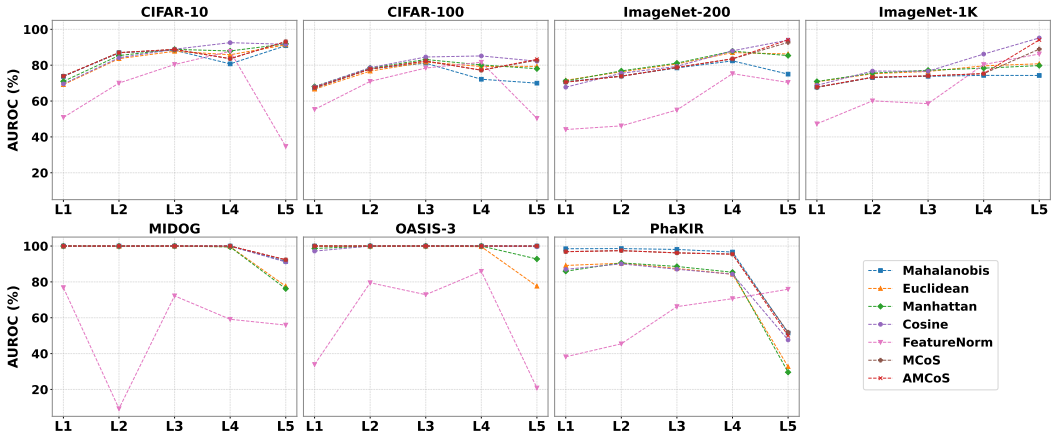


Figure 5: Layer-wise **far-OOD** detection performance. Shallow representations (Layer 1) achieve superior performance in medical domains (MIDOG, OASIS-3, and PhaKIR) but degrade in natural domains (CIFAR-10/100 and ImageNet-200/1k). Conversely, deep representations (Layer 5) exhibit the opposite trend, indicating that optimal OOD discriminability is highly layer-dependent.

A.3 Baseline Performance on far-OOD Samples

Table 7: Comparison of OOD detection methods on **far-OOD** across domains. Our method achieves the best overall performance. **Bold** indicates the best results, and underline denotes the second-best. CIF and IMN denote CIFAR and ImageNet, respectively.

Benchmarks	**OpenOOD**				**OpenMIBOOD**			**Averages for both**		
	CIF-10	CIF-100	IMN-200	IMN-1K	MIDOG	PhaKIR	OASIS-3	$\overline{\text{AUROC}}\uparrow$	$\overline{\text{AUPR}}\uparrow$	$\overline{\text{FPR@95}}\downarrow$
MSP [13]	91.00	77.58	90.20	85.23	80.91	32.51	78.45	76.55	59.79	57.78
ODIN [35]	87.21	79.44	91.89	89.47	85.08	71.82	58.63	80.51	66.55	57.54
MDSEns [17]	91.36	83.15	87.04	77.32	**100.00**	**98.50**	**100.00**	91.05	87.82	26.59
Gram [22]	**96.30**	**86.06**	89.30	90.56	88.27	31.20	**100.00**	83.10	70.00	38.88
KNN [18]	93.10	82.32	93.47	90.18	90.18	37.75	**100.00**	83.86	75.63	38.49
MBM [23]	90.97	82.72	81.97	76.62	**100.00**	96.88	**100.00**	89.88	86.80	27.28
FeatureNorm [25]	87.75	81.64	73.17	86.27	59.14	67.28	72.90	75.45	57.63	62.98
Dynamic [11]	93.26	84.57	**94.71**	94.22	52.26	64.26	**100.00**	83.33	79.32	43.73
S&I [24]	95.46	85.47	87.39	88.71	61.83	92.25	23.88	76.43	63.01	52.45
ProGen [16]	93.25	79.55	91.29	89.32	82.26	71.05	31.10	76.83	60.65	54.07
NCI [19]	91.28	81.37	93.80	**95.53**	90.46	77.89	27.50	79.69	67.09	42.94
ML-AMCoS	94.31	85.71	94.20	94.34	**100.00**	96.15	**100.00**	**94.97**	**92.81**	**18.01**

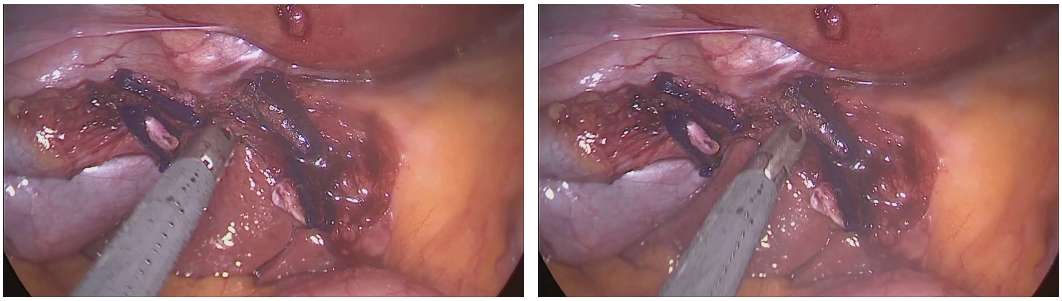
Tab. 7 presents the performance of ML-AMCoS compared to state-of-the-art training-free methods on far-OOD benchmarks. By definition, far-OOD tasks involve samples with significant semantic shifts from the in-distribution (ID) data, making them inherently easier to distinguish than cs-ID or near-OOD scenarios. Consequently, several baseline methods are able to narrow the performance gap; for instance, MDSEns, MBM, and Gram achieve perfect AUROC scores on specific subsets like MIDOG or OASIS-3.

Despite the reduced difficulty of these tasks, ML-AMCoS consistently achieves the best overall performance across both natural and medical image domains. Our method reaches a mean AUROC of 94.97% and a mean AUPR of 92.81%, outperforming the second-best baseline (MDSEns) by 3.92% and 4.99%, respectively. Most notably, our approach demonstrates superior robustness in error reduction, achieving a mean FPR@95 of 18.01%, a substantial 8.58% improvement over the 26.59% achieved by MDSEns.

While some methods excel in specific domains, such as Dynamic [11] on ImageNet-200 or MDSEns on medical datasets, they often suffer from significant performance degradation when transferred to a different domain. In contrast, ML-AMCoS maintains near-perfect or second-best results across

nearly all individual datasets. These results confirm that even in scenarios where OOD cues are more pronounced, our adaptive weighting of feature covariance and layer-wise contributions provides a more robust and generalizable OOD detection.

A.4 Limited Visibility of Smoke in Single-Frame PhaKIR Images



No Smoke Medium Smoke

Figure 6: Subtle smoke cues in PhaKIR. The left frame contains no smoke, while the right frame from the same sequence contains medium smoke. When viewed individually, the smoke is difficult to perceive, illustrating why cs-ID smoke samples provide weak OOD cues in single-frame settings.

The relatively low performance on the cs-ID subset of PhaKIR stems from the inherent difficulty of identifying smoke in static frames. In surgical videos, smoke typically accumulates gradually, and many occurrences, particularly medium levels, are only noticeable when temporal context is available. As shown in Fig. 6, the right frame contains medium smoke, yet without the adjacent clean frame for comparison, the haze is extremely subtle. It produces minimal changes in color, contrast, and edges, making the presence of smoke visually ambiguous even for humans. Consequently, such weak signals are unlikely to induce meaningful shifts in feature activations, leading to cs-ID samples that are nearly indistinguishable from ID images in single-frame analysis.

A.5 Alignment between Layer Weighting w^l and OOD Detection Performance

In this section, we provide extended results for the alignment analysis between our pseudo-OOD weighting strategy and ground-truth OOD detection performance for both cs-ID (Fig. 7) and far-OOD (Fig. 8) tasks. As illustrated in the figures, our corner-cutmixing weighting strategy maintains high ranking precision (NDCG@2) across various OOD types. In the cs-ID task, where OOD samples share semantic labels with the ID data but differ in domain or acquisition style, AMCoS identifies the most discriminative layers with a precision that closely matches the supervised MDSEns baseline. This suggests that the synthetic manifold shifts generated by corner-cutmixing effectively simulate the feature-level disturbances encountered in real-world covariate shifts. While FeatureNorm exhibits competitive alignment on CIFAR-based natural image datasets, its ranking quality collapses in medical imaging benchmarks (MIDOG, OASIS-3, and PhaKIR), often falling below an NDCG@2 of 0.8. In contrast, ML-AMCoS consistently achieves near-perfect alignment on these datasets. This confirms that our approach is significantly more robust to domain-specific feature distributions than Jigsaw-based pseudo-OOD methods, which may rely on spatial structures that are not preserved in medical volumetric or microscopic data. Across both tasks and all seven datasets, the ranking quality remains remarkably stable across the tested range of $\alpha \in [0.1, 0.4]$. This hyperparameter stability indicates that the structural perturbations introduced by corner-cutmixing are not overly sensitive to the interpolation ratio, making it a reliable plug-and-play weighting mechanism for diverse training-free distance metrics, including Mahalanobis, Euclidean, and Manhattan distances.

To evaluate the sensitivity of our approach to the spatial location of perturbations, we report the alignment results for a center-cutmixing variant in Figures 9, 10, and 11. In this configuration, we synthesize pseudo-OOD samples by replacing a central crop of size $\alpha H \times \alpha W$ (with $\alpha \in [0.3, 0.7]$) with a patch from a different class. The empirical results indicate that center-cutmixing achieves ranking performance comparable to the default corner-cutmixing strategy utilized in ML-AMCoS. However, corner-cutmixing demonstrates superior precision on natural image datasets such as CIFAR-

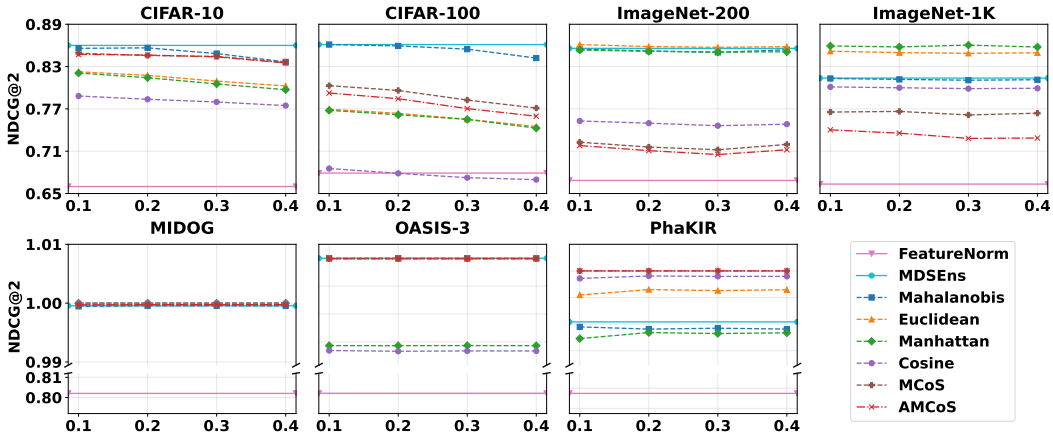


Figure 7: Ranking quality analysis using NDCG@2 across diverse domains and distance metrics on **cs-ID** detection task. Our corner-cutmixing strategy ($\alpha \in [0.1, 0.4]$) is compared against MDSEns (supervised with real near-OOD data) and FeatureNorm (pseudo-OOD using Jigsaw puzzles). Higher values indicate better alignment between layer weights w^l and ground-truth AUROC performance.

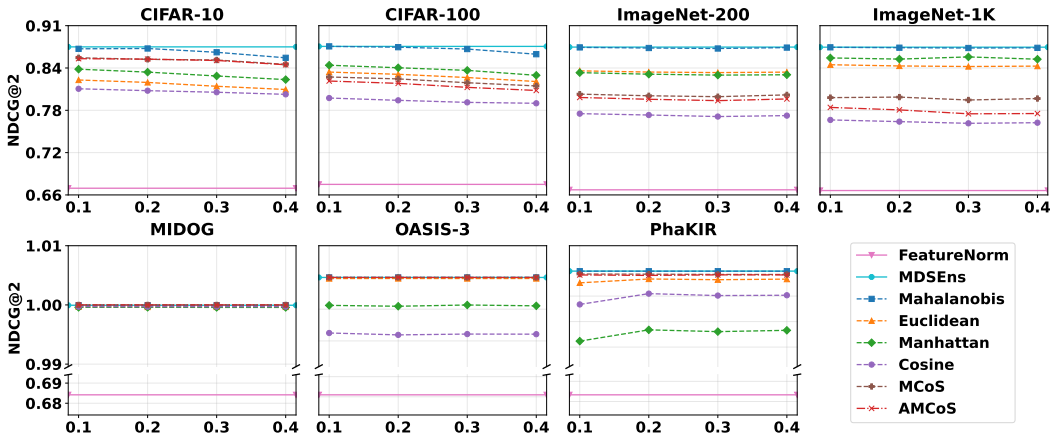


Figure 8: Ranking quality analysis using NDCG@2 across diverse domains and distance metrics on **far-OOD** detection task. Our corner-cutmixing strategy ($\alpha \in [0.1, 0.4]$) is compared against MDSEns (supervised with real near-OOD data) and FeatureNorm (pseudo-OOD using Jigsaw puzzles). Higher values indicate better alignment between layer weights w^l and ground-truth AUROC performance.

10 and CIFAR-100. Specifically, the ranking quality of the Mahalanobis distance under corner-cutmixing aligns more closely with the supervised MDSEns baseline than the center-cutmixing alternative. This discrepancy suggests that in low-resolution datasets like CIFAR-10 and CIFAR-100, central perturbations are more likely to obscure the primary object of interest, which typically occupies the center of the frame. This results in the generation of a mislabeled sample rather than a representative pseudo-OOD shift.

A.6 Ablation Study on far-OOD

In this section, we extend our ablation analysis to far-OOD scenarios to evaluate the consistency of ML-AMCoS's components when semantic shifts are more pronounced. The results are summarized in Tab. 8. Similar to the observations in the near-OOD experiments, performance on far-OOD tasks is generally stable across natural image datasets (CIFAR and ImageNet) as the number of aggregated layers increases. However, the catastrophic failure on the PhaKIR benchmark persists when the selection extends beyond the top-2 layers, with AUROC scores decreasing by over 36%. This confirms that non-discriminative layers can introduce substantial noise that degrades the detection of domain-

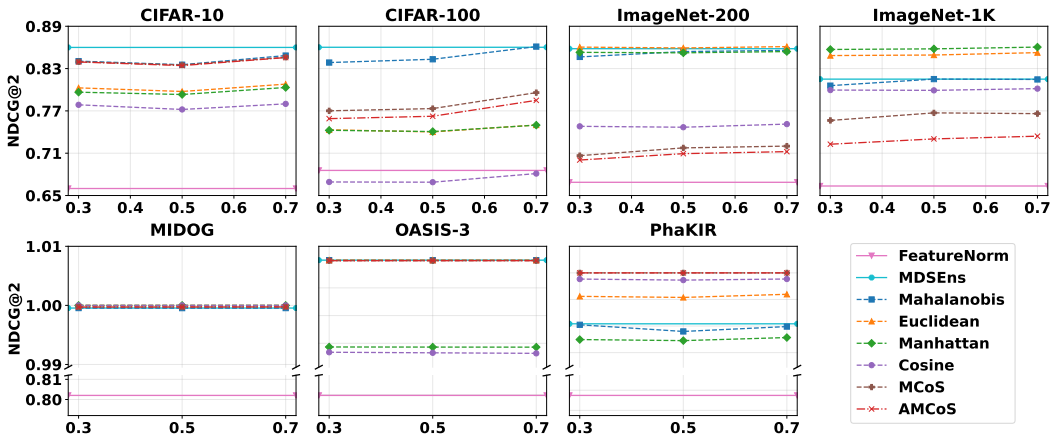


Figure 9: Ranking quality analysis using NDCG@2 across diverse domains and distance metrics on **cs-ID** detection task under center-cutmixing strategy ($\alpha \in [0.3, 0.7]$). Higher values indicate better alignment between layer weights w^l and ground-truth AUROC performance.

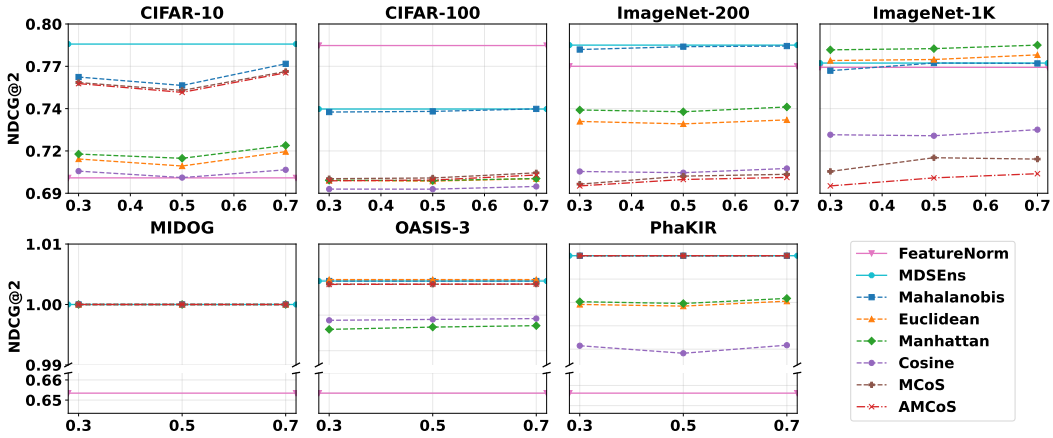


Figure 10: Ranking quality analysis using NDCG@2 across diverse domains and distance metrics on **near-OOD** detection task under center-cutmixing strategy ($\alpha \in [0.3, 0.7]$). Higher values indicate better alignment between layer weights w^l and ground-truth AUROC performance.

specific anomalies, even in far-OOD settings. Furthermore, the removal of performance-driven weighting (w/o w^l) leads to marginal performance gains on simpler natural image datasets but causes a severe collapse on PhaKIR. This suggests that while uniform averaging might suffice for broad semantic shifts, it lacks the precision required for specialized medical domains. Finally, the removal of adaptive covariance (w/o λ) results in a significant performance drop on ImageNet-1K (-5.32%), the most complex far-OOD benchmark in our suite. This decline underscores the importance of mitigating covariance noise in layers with a high trace, where raw statistical estimates become less reliable.

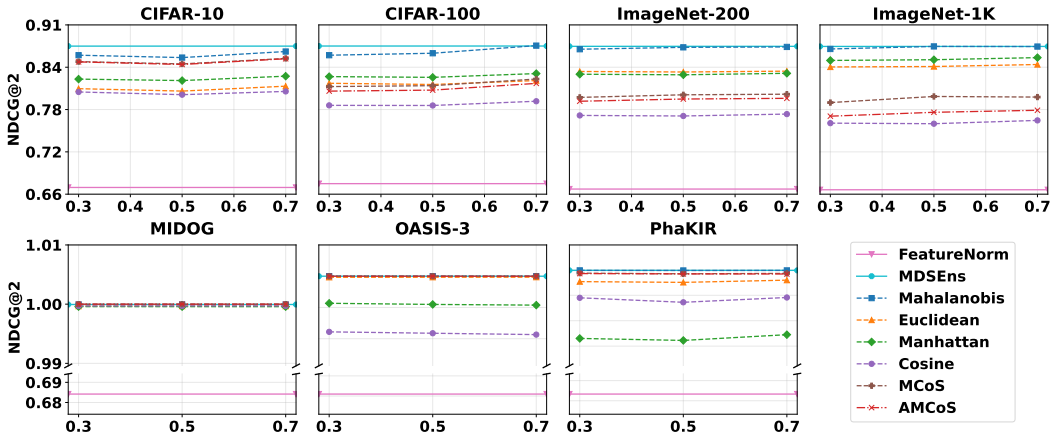


Figure 11: Ranking quality analysis using NDCG@2 across diverse domains and distance metrics on **far-OOD** detection task under center-cutmixing strategy ($\alpha \in [0.3, 0.7]$). Higher values indicate better alignment between layer weights w^l and ground-truth AUROC performance.

Table 8: Ablation analysis of ML-AMCoS configurations. We report AUROC scores on **far-OOD** across all seven datasets. The default setting utilized in the main results (top-2) is highlighted with an asterisk (*). For all other settings, performance variations relative to the default setting are indicated in the format $\text{AUROC}_{\Delta\text{AUROC}}$.

	CIF-10	CIF-100	IMN-200	IMN-1K	MIDOG	PhaKIR	OASIS-3
ML-AMCoS (top-1)	$93.16_{(-1.15)}$	$83.79_{(-1.92)}$	$92.73_{(-1.47)}$	$94.12_{(-0.22)}$	100.00	$95.47_{(-0.68)}$	100.00
ML-AMCoS (top-2)*	94.31	85.71	94.20	94.34	100.00	96.15	100.00
ML-AMCoS (top-3)	$94.57_{(+0.26)}$	$86.60_{(+0.89)}$	$95.46_{(+1.26)}$	$94.29_{(-0.05)}$	100.00	$57.97_{(-38.18)}$	100.00
ML-AMCoS (top-4)	$94.79_{(+0.48)}$	$86.92_{(+1.21)}$	$95.48_{(+1.28)}$	$94.13_{(-0.21)}$	$99.92_{(-0.08)}$	$58.64_{(-37.51)}$	100.00
ML-AMCoS (top-5)	$94.83_{(+0.52)}$	$87.06_{(+1.35)}$	$95.50_{(+1.30)}$	$94.08_{(-0.26)}$	$99.92_{(-0.08)}$	$59.54_{(-36.61)}$	100.00
ML-AMCoS (w/o w^l)	$95.14_{(+0.83)}$	$87.84_{(+2.13)}$	$95.40_{(+1.20)}$	$93.28_{(-1.06)}$	100.00	$58.70_{(-37.45)}$	100.00
ML-AMCoS (w/o λ)	$94.24_{(-0.07)}$	$85.58_{(-0.13)}$	$92.88_{(-1.32)}$	$89.02_{(-5.32)}$	100.00	$96.17_{(+0.02)}$	100.00

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: We provide a detailed explanation of our methodology in Sec. 4 and demonstrate empirical improvements in AUROC within Sec. 5.

   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: We discuss the limitations of our approach in Sec. 5 and Conclusion.

   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: [N/A]

Justification: No theory assumptions or proofs in this paper.

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: We provide a comprehensive description of our methodology in Sec. 4. Furthermore, all datasets used in our experiments are publicly available and cited in Sec. 5.

   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: All datasets used in our experiments are publicly available and cited in Sec. 5.

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: We utilize publicly available benchmarks for all experiments and adhere to the standard parameter settings established by these benchmarks, as detailed in Sec. 5 and Appx. A.1.

   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: Due to the significant scale of the evaluation datasets and the observed stability of the results across multiple runs, we report mean performance metrics. We found that the variance was negligible and did not qualitatively alter the comparative rankings of the methods.

   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: We list the hardware usage in Appx. A.1.

   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: All datasets used in this paper are publicly available.

   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: [N/A]

    Justification: This work focuses on fundamental research in training-free out-of-distribution detection. While it improves model reliability in domains like medical imaging (public datasets), it does not present immediate societal impacts beyond general improvements in machine learning robustness.

    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: [N/A]

    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 used are properly cited. We adhere to the specific licenses and terms of use for each, including the formal application processes required for restricted-access benchmarks.

    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: [N/A]

    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: [N/A]

    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: [N/A]

    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: [No]

    Justification: LLMs were used solely for grammatical revision and stylistic refinement of the manuscript; they did not contribute to the core methodology, data analysis, or scientific results.

    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.