Why Feature Magnitude Deceives OOD Detectors: An Angular Separation Perspective

Hanlin Li[1], Jing Ma[1], Zehang Wei[1], Jiamin Yan[1], and Xiang Xiang[1,2⋆]

[1] Huazhong University of Science and Technology, Wuhan, China
[2] Pengcheng Laboratory, Shenzhen, China
⋆ xex@hust.edu.cn

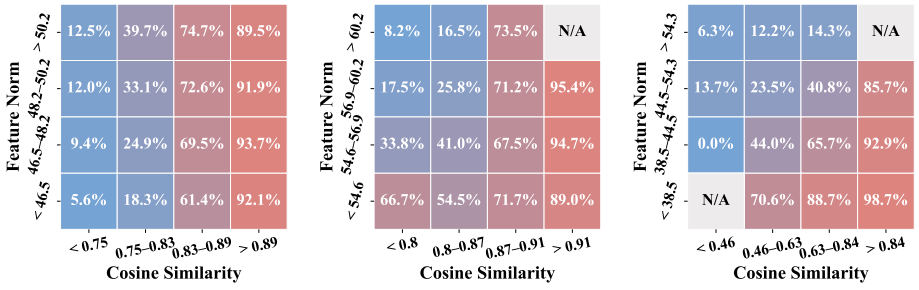
Abstract. Out-of-Distribution (OOD) detection is a critical safety requirement for deploying deep neural networks in open-world environments. While recent advances increasingly rely on more computationally intensive training methods involving synthetic outliers, contrastive objectives, or specialized loss functions, their gains often come with substantial computational overhead and implementation complexity. In this work, we revisit the fundamentals of OOD detection and uncover a key flaw in common distance-based detectors: sensitivity to feature magnitude. We show that low-norm OOD samples can appear closer to in-distribution (ID) class centroids than actual ID samples, evading detection. To this end, we introduce **Angular Separation Learning (ASL)**, a simple and highly effective strategy that applies ℓ_2-normalization to features before the final classification layer. This modification compels the network to optimize for angular separation, achieving robust feature learning without additional regularization mechanisms, synthetic samples, or costly negative mining. Through extensive experiments on diverse benchmarks, we demonstrate that ASL not only matches but often surpasses state-of-the-art methods, especially in challenging near-OOD scenarios, while maintaining training efficiency. Our results indicate that a minimalist rethink of standard training can achieve superior OOD performance, prompting a re-evaluation of the complexity-to-performance trade-off in OOD detection. Codes are available at `https://github.com/HAIV-Lab/ASL`.

Keywords: Out-of-Distribution Detection · Feature Normalization

1 Introduction

The deployment of deep learning models into real-world, safety-critical applications represents a major frontier for artificial intelligence. From autonomous vehicles navigating unpredictable roadways [21] to AI-powered systems aiding in medical diagnosis [61], models are increasingly tasked with making high-stakes decisions in dynamic, open-world environments [19, 39]. However, a fundamental gap exists between the controlled conditions of training and the unbounded nature of reality. Models trained under a "closed-world" assumption often fail

⋆ Corresponding author.



(a) Raw Features **(b)** Normalized Features **(c)** ASL Trained

Fig. 1: Visualization of ID vs. OOD separability in the Mahalanobis space. The two-dimensional space is defined by the feature's ℓ_2-norm and its cosine similarity to the nearest class mean. Space is partitioned into 4×4 regions using independent thresholds computed for each dimension, and each region indicates the proportion of in-distribution samples within that area. ASL enhances angular separation and concentration, resulting in more discriminative ID/OOD distributions and improved detection.

catastrophically when encountering inputs that deviate from their training distribution, known as out-of-distribution (OOD) samples [2]. A self-driving car may misinterpret an unfamiliar object on the road, or a medical diagnostic tool may produce a dangerously overconfident prediction for a rare disease. This failure to recognize and appropriately handle novel inputs is a primary barrier to safe and reliable deployment. Consequently, the ability to detect OOD samples is not merely a desirable property but a critical safety requirement for building trustworthy AI systems that can robustly "know when they don't know".

The core challenge of OOD detection is to distinguish between in-distribution (ID) and OOD inputs using a model trained on ID data. Initial and ongoing research has produced a rich landscape of post-hoc methods that operate on pre-trained models without altering their weights. These approaches leverage various signals from the network, from simple uncertainty scores derived from the model's output logits like MSP [18] and Energy scores [29], to more sophisticated techniques that analyze the network's internal state. Researchers have successfully utilized backpropagation gradients [20] and manipulated intermediate activation patterns [9, 45] to amplify the distinction between ID and OOD samples. Among this diverse array of techniques, methods that analyze the geometric properties of the feature space have proven particularly powerful. Specifically, distance-based detectors, such as the Mahalanobis Distance Score (MDS) [25], have demonstrated strong empirical performance, particularly under the closed-world assumption or with far-OOD samples.

While these post-hoc methods are valuable, their effectiveness is fundamentally capped by the representations they inherit. A standard cross-entropy objective trains a model to be discriminative, learning only the minimal features necessary to separate the known ID classes rather than modeling the underlying data distribution [51]. This can result in a feature space where ID classes are not well-separated and OOD samples are incorrectly projected into regions of high

confidence [34]. This realization has motivated a crucial shift towards training-based approaches that explicitly sculpt the feature manifold for OOD robustness. The core advantage of these methods is the ability to learn representations with superior geometric properties, such as increased intra-class compactness and inter-class separability, thereby creating a more structured and reliable feature space [55]. While effective, these methods often incur substantial computational overhead, hyperparameter sensitivity, and implementation intricacy. More importantly, Fig. 2 suggests that this complexity does not always reliably translate into superior performance, especially for the challenging near-OOD case, raising a fundamental question: is such complexity always justified?

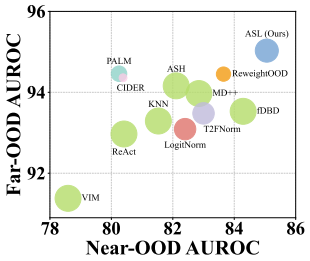


Fig. 2: OOD detection performance on ImageNet-200. Larger point size represents higher training speed, and green points indicate post-hoc methods using model trained on cross-entropy.

In stark contrast to this trend, we analyze a key vulnerability in Mahalanobis distance-based detectors: their extreme sensitivity to the scale (ℓ_2-norm) of features. The Mahalanobis distance calculation conflates a feature's magnitude with its directional information. Our analysis reveals that because ID and OOD samples often exhibit distinct norm distributions, the scoring mechanism is disrupted. This creates a pathological failure mode: a low-norm OOD sample, despite its potentially large angular deviation, can yield a smaller Mahalanobis distance than a genuine high-norm ID sample, thereby evading detection. To address this fundamental problem, we introduce Angular Separation Learning (ASL), which forces the model to learn representations optimized for angular separation, rather than relying on feature magnitude. Surprisingly, this simple modification alone is sufficient to significantly enhance intra-class compactness, which is evidenced by the tighter clustering of ID samples in regions of high cosine similarity, and this angular concentration leaves well-defined low-density regions between classes, as shown in Fig. 1. In summary, our contributions are as follows:

- We provide a detailed analysis of a specific flaw in Mahalanobis-based detection in Mahalanobis-based OOD detection, revealing that its root cause lies in the feature space learned by standard cross-entropy training.
- We propose ASL, a training framework that applies ℓ_2-normalization to the feature representations before the cross-entropy loss. This simple modification optimizes the feature space for angular discriminability, creating robust representations for OOD detection.
- We conduct extensive experiments on multiple OOD benchmarks, demonstrating that our method significantly outperforms existing state-of-the-art approaches, particularly in challenging near-OOD scenarios.

2 Related Work

Distance-Based Post-Hoc OOD Detection. These methods analyze the geometric properties of a pre-trained model's feature space. A prominent line of work builds upon the Mahalanobis Distance Score (MDS) [25], which models classes as conditional Gaussians. Refinements include VIM [53], which incorporates a residual vector, and RMDS [43], which mitigates the influence of non-discriminative features by subtracting a global background distance from the class-conditional one. fDBD [27] approximates the distance to the decision boundary. MD++ [35] addresses sensitivity to feature norms directly by applying normalization to features before distance computation. Non-parametric alternatives avoid distributional assumptions, such as the k-NN detector [46], and its extension NNGuide [38] that uses k-NN to guide confidence scores.

Angular and Neural Collapse Based Methods. Recent methods continue to explore richer geometric structures beyond Euclidean distance. A prominent line of work builds upon the phenomenon of neural collapse and angular geometry for post-hoc detection. [6] propose ORA, a scale-invariant angular score between features and decision boundaries relative to ID mean. In parallel, the phenomenon of neural collapse has inspired new detectors: NCI [28] combines centered angular proximity to weight vectors with feature norm filtering, whereas NECO [1] leverages Neural Collapse, computing PCA-projected norm ratio with MaxLogit calibration. In parallel, another thread of research investigates training-based angular learning. NormFace [52] explores joint normalization of features and weights on a hypersphere, which effectively fixes the softmax temperature. SphereFace [30] introduces a multiplicative angular margin loss for face recognition to enforce larger inter-class separations on a hypersphere, while [48] propose a hyperparameter-free cosine-similarity classifier for OOD detection that learns a scaling parameter via an additional branch.

OOD Detection via Contrastive Learning. Training-based methods for OOD detection often leverage contrastive learning to learn more separable feature representations by enhancing intra-class compactness and inter-class separation. They can be categorized based on their source of negative samples for the contrastive objective. Many use other ID samples as negatives, often building on frameworks like SimCLR [4] and Supervised Contrastive Learning (SupCon) [22]. For instance, SSD [44] uses self-supervised contrastive pre-training followed by fine-tuning, while CIDER [34] explicitly optimizes mutual information to tighten class clusters. PALM [31] represents each class with a mixture of prototypes to capture complex intra-class structures. More recently, ReweightOOD [42] dynamically reweights negative pairs based on embedding similarity to further reduce intra-class variance. Another strategy explicitly defines the decision boundary by generating pseudo-OOD samples to serve as negative examples. These methods contrast in-distribution (positive) samples against the synthesized (negative) outliers. VOS [11] synthesizes virtual outliers from low-likelihood regions of class-conditional feature distributions, while NPOS [47] adopts a non-parametric approach by perturbing boundary ID samples. HamOS [26] uses Hamiltonian Monte Carlo to efficiently generate diverse synthetic outliers.

VLM-based OOD Detection. Recently, Vision-Language Models (VLMs) like CLIP [40] have driven significant advancements in OOD detection. Early zero-shot frameworks such as MCM [33] rely on aligning visual features with textual concepts (ID class names) via softmax-scaled cosine similarity. To further enrich ID representations, Dual-Pattern Matching (DPM) [59] integrates both aggregated visual prototypes and textual patterns, establishing a stricter multimodal alignment. Moreover, recent prompt-guided approaches [15] utilize Large Language Models to synthesize class-specific positive and negative prompts, explicitly modeling inter-class boundaries and transferring this semantic supervision to the visual branch. While these methods achieve robust performance, they inherently depend on dual encoders and massive image-text pre-training. In contrast, our ASL focuses on a complementary setting: improving unimodal, lightweight vision models trained from scratch.

Normalization. A growing body of work has highlighted the benefits of normalization for out-of-distribution (OOD) detection. These efforts can be broadly grouped into two main categories. The first leverages the norm itself as a discriminative signal: for instance, [58] incorporate feature norms directly into their scoring function, while [37] identify the feature norm as a class-agnostic confidence measure. However, under standard cross-entropy training, we observe that this separability is often compromised. Models tend to exploit optimization shortcuts by inflating feature magnitudes, leading to significant overlap between ID and OOD norm distributions (Supp. Fig. 9, Fig. 10). This indicates that reliable norm separability is not an intrinsic property of unconstrained training, but rather necessitates explicit geometric constraints to be effectively realized. The second category applies normalization during training primarily as a form of regularization, early approaches like LogitNorm [54] focused on regularizing the model's output space to mitigate overconfidence. Building on this, T2FNorm [41] demonstrated that normalizing features during training could improve logit separation at test time. Concurrently, from a theoretical standpoint, [12] established a crucial link between ℓ_2-normalization, earlier neural collapse, and improved OOD performance. Although these studies confirm the utility of normalization, which acts either as a scoring cue or a regularizer, they largely treat it as an auxiliary mechanism. In contrast, we propose a unified perspective that elevates ℓ_2-normalization to the core of the learning objective. We demonstrate that it performs contrastive learning in angular space, offering a streamlined and highly effective alternative to complex OOD detection approaches.

3 Method

3.1 Preliminaries

OOD Detection Setup. Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{1, \ldots, K\}$ be the label space for a K-class classification task. We are given an in-distribution (ID) training dataset $\mathcal{D}_{in} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ sampled from a joint distribution $P_{in}(\boldsymbol{x}, y)$. A deep neural network $f(\cdot; \theta)$ is trained on $\mathcal{D}_{in}$, which can be decomposed into

an encoder $f(\cdot) : \mathcal{X} \to \mathbb{R}^d$ that maps an input $\boldsymbol{x}$ to a d-dimensional feature embedding, and a classifier that outputs a class prediction.

The goal of out-of-distribution (OOD) detection is to distinguish ID samples (from P_{in}) from OOD samples (from an unknown $P_{out}(\boldsymbol{x})$ with disjoint label space). An ideal scoring function should yield higher scores for ID inputs than for OOD inputs, i.e., $S(\boldsymbol{x}_{in}) > S(\boldsymbol{x}_{out})$ for $\boldsymbol{x}_{in} \sim P_{in}$ and $\boldsymbol{x}_{out} \sim P_{out}$.

A widely used baseline is the **Mahalanobis Distance Score (MDS)** [25]. It models the feature distribution of each class c as a Gaussian $\mathcal{N}(\boldsymbol{\mu}_c, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_c$ is the class mean and $\boldsymbol{\Sigma}$ is a shared covariance matrix, both estimated from $\mathcal{D}_{in}$. The OOD score for a test sample $\boldsymbol{x}$ and its feature $f(\boldsymbol{x})$ is defined as:

Table 1: Comparison of OOD detection performance (FPR95↓ and AUROC↑). FeatCos refers to using only the cosine similarity between features and class means as the detection score. FeatCos† uses features from an ASL trained model.

ID Dataset	Method	Near-OOD		Far-OOD	
		FP↓	AU↑	FP↓	AU↑
CIFAR -10	MSP	45.8	88.3	30.0	91.4
	MDS	49.2	85.2	33.3	89.0
	MD++	40.2	88.8	30.6	91.0
	FeatCos	38.9	89.4	29.6	91.3
	FeatCos†	30.5	92.0	16.0	96.4
CIFAR -100	MSP	55.9	80.0	58.5	78.1
	MDS	83.0	59.4	73.5	68.3
	MD++	67.8	74.8	51.9	82.6
	FeatCos	59.0	80.3	54.6	81.8
	FeatCos†	57.1	80.3	50.4	82.7

$$S_{MDS}(\boldsymbol{x}) = -\min_{c \in \mathcal{Y}} \left((f(\boldsymbol{x}) - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}^{-1} (f(\boldsymbol{x}) - \boldsymbol{\mu}_c) \right),$$

this score assumes that OOD samples lie farther from all class centroids than ID samples.

3.2 Motivating Observations

The Mechanics of MDS Failure. As shown in Tab. 1, MDS and its variant MD++ [35] are often outperformed by the simple cosine similarity score.

To understand why MDS fails to deliver satisfactory performance, we can decompose the Mahalanobis distance calculation. For a given feature vector $f(x)$, the distance to a class-c prototype $\hat{\mu}_c$ is:

$$dist_c(\boldsymbol{x}) = \underbrace{f(\boldsymbol{x})^\top \hat{\Sigma}^{-1} f(\boldsymbol{x})}_{\text{Sample Quadratic Term (SQT)}} - 2 \underbrace{f(\boldsymbol{x})^\top \hat{\Sigma}^{-1} \hat{\mu}_c}_{\text{Cross-Sample Term (CST)}} + \underbrace{\hat{\mu}_c^\top \hat{\Sigma}^{-1} \hat{\mu}_c}_{\text{Class Bias Term (CBT)}}. \tag{1}$$

Theoretically, minimizing the cross-entropy loss should align with minimizing Eq. (1) by simultaneously reducing the feature norm and increasing the alignment with the class mean. However, in practice, we observe a significant optimization bias during standard training. The model tends to exploit a 'lazy'

shortcut by inflating feature norms to rapidly minimize the loss, rather than investing in the more difficult task of enforcing strict intra-class compactness. This unbalanced training dynamic results in a feature space where large variance in magnitude acts as a confounding factor, making the final MDS score highly sensitive to feature norms rather than reflecting true semantic distance.

To formalize this analysis, we can simplify the expression by performing a whitening transformation. Let $g(x) = \hat{\Sigma}^{-1/2} f(x)$ and $\hat{u}_c = \hat{\Sigma}^{-1/2} \hat{\mu}_c$ be the feature and class mean in the whitened space (also known as the Mahalanobis space), respectively. Ignoring the constant term CBT and letting $\theta(x, y)$ represent the angle between two vectors, then for category c the Mahalanobis distance score can be written as:

$$S_c(\boldsymbol{x}) \approx -\|g(\boldsymbol{x})\|_2^2 + 2\|g(\boldsymbol{x})\|_2 \|\hat{u}_c\|_2 \cos(\theta(g(\boldsymbol{x}), \hat{u}_c)). \tag{2}$$

Empirically, the norm of a typical ID feature, $\|g(\boldsymbol{x})\|_{ID}$, is close to the norm of its corresponding class mean, $\|\hat{u}_c\|_2$. Since $\cos(\theta) \leq 1$, the score is maximized at a norm value that is less than or equal to the typical ID feature norm. This implies that for any feature $g(x)$ whose norm is greater than the optimal value (i.e., $\|g(x)\|_2 > \|\hat{u}_c\|_2 \cos\theta$), increasing its norm will decrease the score, correctly identifying it as an OOD sample. However, it also exposes a critical failure mode: an OOD sample with an anomalously small norm can achieve a deceptively high score.

We demonstrate this pathology with an empirical example (see Fig. 1a), where the whitened class mean norm is $\|\hat{u}_c\|_2 = 48$: 1) A **typical ID sample** with high similarity ($\cos\theta = 0.825$) and representative norm ($\|g(x)\|_2 = 49$) attains a score of **1480**. 2) A **near OOD sample** with equally high similarity ($\cos\theta = 0.82$) but smaller norm ($\|g(x)\|_2 = 47$) achieves a higher score of **1491**. 3) A more **dissimilar OOD sample** ($\cos\theta = 0.81$) with an even smaller norm ($\|g(x)\|_2 = 40$) receives the highest score (**1510**). Notably, despite lower alignment with the class prototype, its significantly reduced norm yields a falsely confident score.

Work such as MD++ has observed that feature normalization can mitigate the interference of low-norm OOD samples when using MDS. However, as shown in Fig. 1b, while normalization increases the proportion of ID samples in lower-norm regions to some extent, it fails to improve angular distribution. As a result, the normalized model still exhibits near-OOD performance comparable to, or even worse than, the simple cosine similarity score.

This analysis uncovers a counter-intuitive reality: distance-based scoring functions, such as MDS, are structurally vulnerable to confounding feature magnitudes, allowing low-norm OOD samples with poor semantic similarity to appear more in-distribution than genuine ID samples. This pathology exposes a fundamental failure mode in relying on Euclidean distance spaces optimized by standard cross-entropy. This motivates our central research question: *Can we go further to learn a feature geometry that is intrinsically invariant to norm variations and emphasizes pure angular separation to universally improve OOD detection?*

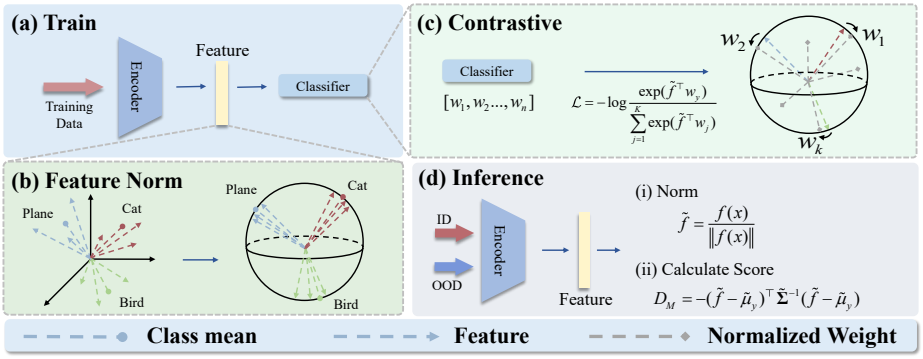


Fig. 3: Pipeline of the proposed ASL framework. During training, features are ℓ_2-normalized before cross-entropy loss. This encourages angular separation between classes without explicit contrastive learning.

3.3 ASL: Training for Optimal Angular Separation

Our diagnosis of Mahalanobis failures bridges the gap between the vulnerabilities of distance-based scoring and the necessity for robust representation learning. As highlighted by the results in Tab. 1, while feature norm-sensitive scores are heavily disrupted, utilizing a purely angular metric (Cosine Similarity) achieves highly competitive performance, particularly in challenging near-OOD scenarios.

These observations directly inspire our core idea: given the criticality of angular information for distinguishing OOD samples, rather than passively dealing with the interference caused by norm variations during OOD detection, we should proactively enforce the model during training to learn a feature space optimized purely for angular separation, thereby generating feature representations that are insensitive to norm variations and inherently robust.

To achieve this goal, we propose **Angular Separation Learning (ASL)**. ASL is built upon a key insight: when the standard cross-entropy loss is applied to ℓ_2-normalized features, it performs contrastive learning. This simple operation forces the model to focus exclusively on optimizing the angular separation between classes. We will show that this method is equivalent to a form of contrastive learning while avoiding the computational overhead of explicit pair mining or complex loss functions, and this simplicity is also key to the high training efficiency of ASL. Specifically, it only adds a normalization step after feature extraction, along with a further simplification of removing the bias term from the linear layer. This results in a computational cost that is nearly on par with standard cross-entropy training, while achieving lower training.

Let the feature extractor be $f(\cdot)$ and the final linear classifier have weights $\{w_j\}_{j=1}^{K}$, which we treat as learnable prototypes for each class. We first ℓ_2-normalize the feature vector: $\tilde{f} = s \cdot f(x)/||f(x)||_2$. Here, s is a scaling factor introduced to counteract the reduced gradient magnitudes resulting from ℓ_2-normalization, and it is typically set to 10. The logit for class j then becomes

$w_j^\top \tilde{f} = ||w_j||_2 \cos(\theta_j)$, where θ_j is the angle between the feature vector $\tilde{f}$ and the class prototype w_j. To prevent the model from simply increasing the weight norms $||w_j||_2$ to minimize the loss, we employ standard weight decay. The final training objective for ASL is:

$$\mathcal{L}_{\text{ASL}} = \mathcal{L}_{\text{CE}}(\mathbf{W}^\top \tilde{f}, y) + \lambda ||\mathbf{W}||_F^2,$$

where $\tilde{f} = s \cdot f(x)/||f(x)||_2$ is the normalized feature vector, y is the class label of sample x, $\mathbf{W}$ is the weight matrix of the final classifier, λ is the weight decay hyperparameter, and $||\cdot||_F$ denotes the Frobenius norm. This simple yet effective objective compels the model to learn a feature space where classes are distinguished purely by their angular position on the hypersphere.

The following proposition formalizes the properties of this approach.

Proposition 1 (Angular Separation Learning). *Let $\tilde{f} = s \cdot f(x)/\|f(x)\|_2$ be ℓ_2-normalized features and $\mathbf{W} = [w_1, \ldots, w_K] \in \mathbb{R}^{d\times K}$ a prototype matrix. Minimizing the loss*

$$\mathcal{L}_{ASL} = -\log \frac{\exp(\tilde{f}^\top w_y)}{\sum_{j=1}^{K} \exp(\tilde{f}^\top w_j)} + \lambda \|\mathbf{W}\|_F^2, \quad \lambda > 0, \tag{3}$$

induces a hyperspherical feature geometry with the following properties: (i) It minimizes the angular distance θ_y between $\tilde{f}$ and w_y while maximizing the angular distances θ_j $(j \neq y)$ between $\tilde{f}$ and all other prototypes w_j; (ii) At convergence, $|w_j|_2 \leq \frac{s}{2\lambda}$ for all j.

As formally established in Proposition 1, the optimization process performs prototype-based contrastive learning in the angular space, simultaneously minimizing the angle to the correct class prototype while maximizing the angles to all incorrect ones, and the weight decay λ imposes an upper bound on the norms of the classifier weights, which in turn function as an inverse temperature controlling the sharpness of the softmax distribution, while this temperature regulation is dynamic (unlike fixed-temperature in contrastive learning), allowing it to adapt from coarse-grained early training to fine-grained late-stage optimization. See Appendix Sec. 5.2 for a more complete discussion.

Proposition 2 (OOD Separability Lower Bound of ASL). *Let $S_P = \{(x_i, y_i)\}_{i=1}^n$ be i.i.d. samples from the in-distribution $P(X,Y)$, $S_{cs} = \{(x_j, y_j)\}_{j=1}^m$ be i.i.d. samples from the covariate shift distribution $Q_{cs}(X,Y)$, and $S_{ood} = \{x_k\}_{k=1}^m$ be i.i.d. samples from the out-of-distribution (OOD) distribution $Q_{ood}(X)$ (with label space disjoint from P). Define $\Delta \triangleq d_{\mathcal{F}_{ASL}}(P, Q_{ood}) - d_{\mathcal{F}_{ASL}}(P, Q_{cs})$. For any confidence level $\delta \in (0,1)$, the following holds with probability at least $1-\delta$:*

$$\begin{aligned} \Delta \geq & \frac{s}{2\lambda} \left(\|\hat{\mu}_P - \hat{\mu}_{ood}\|_2 - \|\hat{\mu}_P - \hat{\mu}_{cs}\|_2 \right) \\ & - \frac{2s^2}{\lambda} \left[1 + \sqrt{\frac{\log(8/\delta)}{2}} \right] \left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{m}} \right), \end{aligned} \tag{4}$$

where $n = |S_P|$ *denotes the size of the in-distribution sample set, and* $m = |S_{cs}| = |S_{ood}|$ *represents the size of the covariate shift/OOD sample sets* [1], $\hat{\mu}_D = \frac{1}{|S_D|}\sum_{x \in S_D} \tilde{f}(x)$ *is the empirical feature mean of the sample set* S_D *for distribution* D [2], $d_{\mathcal{F}_{ASL}}(P, Q)$ *denotes the Integral Probability Metric (IPM) induced by the ASL-sensitive function class* $\mathcal{F}_{ASL}$ *that measures the separability between distributions* P *and* Q.

Building upon the bounded weight space, Proposition 2 establishes a theoretical lower bound (Δ in Eq. (4)) for the feature-space separability between ID and OOD samples. When examined through the lens of optimization dynamics, this bound offers insight into why increasing λ can improve near-OOD detection. For challenging near-OOD samples, increasing λ provides two potential benefits. From a statistical perspective, it restricts the hypothesis space more tightly, which can help suppress the negative statistical penalty (the second term scaling with $1/\lambda$). From an optimization perspective (as suggested by Proposition 1), a larger λ encourages the network to learn finer-grained, more discriminative features. This dynamic may actively push the empirical OOD mean ($\hat{\mu}_{ood}$) away from the ID mean ($\hat{\mu}_P$), effectively enlarging the true geometric distance. Consequently, the signal term can be expanded while the noise penalty is suppressed, contributing to a tighter lower bound. For far-OOD samples, the inherent semantic discrepancy already ensures a relatively large initial distance. Since the representations are optimized within a bounded hyperspherical space (e.g., angular separation), the potential for further distance expansion is geometrically limited. Thus, the dynamic effect of λ tends to have limited impact, which may explain why far-OOD detection generally maintains high separability and appears less dependent on λ adjustments.

4 Experiments

4.1 Experimental Setup

Datasets and Models. All experiments follow the standard benchmark protocols in OpenOOD [57] for fair comparison. We adopt four main in-distribution (ID) datasets: CIFAR-10, CIFAR-100 [23], ImageNet-200, and ImageNet-1K [7].For CIFAR-10/100, near-OOD datasets include each other and Tiny ImageNet (TIN) [24], while far-OOD covers MNIST [8], SVHN [36], Texture [5], and Places365 [60]. For ImageNet-1K/200, near-OOD evaluation employs SSB-hard [51] and NINCO [3], and far-OOD uses iNaturalist [49], Texture, and OpenImage-O

[1] Unified for derivational simplicity; if the two sample sizes are different ($m_1 = |S_{cs}|, m_2 = |S_{ood}|$), we only need to replace $\frac{1}{\sqrt{m}}$ in the lower bound with $\frac{1}{\sqrt{m_1}}$ and $\frac{1}{\sqrt{m_2}}$, the proof process is completely consistent.

[2] The feature extractor $\tilde{f}(\cdot)$ is pre-trained on an independent in-distribution training set disjoint from S_P, S_{cs}, S_{ood}, which serve only as test samples for mean estimation in this proof.

Table 2: OOD detection performance on ImageNet-200 using a ResNet-18 model trained from scratch. CE represents model trained with cross-entropy. **Best results** are presented in bold, and second-best results are presented in underline.

	Near-OOD						Far-OOD								
	SSB-hard		NINCO		Avg.		iNaturalist		Textures		OpenImage-O		Avg.		
Method	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	Acc.↑
CE	72.5	78.8	46.1	86.9	59.3	82.8	19.1	95.9	24.3	94.6	29.8	91.4	24.4	94.0	86.3
CIDER	71.0	75.5	45.3	85.3	58.1	80.4	20.0	94.4	16.8	97.1	27.7	91.6	21.5	94.4	84.2
LogitNorm	66.7	78.4	46.6	86.4	56.7	82.4	15.7	96.1	31.7	92.3	31.0	90.9	26.1	93.1	86.4
PALM	70.1	75.7	50.3	84.8	60.2	80.2	21.9	94.3	**12.7**	**97.8**	29.3	91.3	21.3	94.5	83.8
ReweightOOD	67.2	80.5	41.8	86.8	54.5	83.7	18.2	95.0	19.5	96.4	27.8	91.9	21.9	94.5	84.0
T2FNorm	66.1	79.2	46.2	86.8	56.1	83.0	**13.7**	**96.8**	33.7	91.8	28.5	91.8	25.3	93.5	86.5
ASL (Ours)	**65.7**	**81.5**	**39.1**	**88.6**	**52.4**	**85.1**	15.2	96.5	21.6	95.4	**24.4**	**93.2**	**20.4**	**95.0**	**86.8**

Table 3: OOD detection performance on CIFAR-10 and CIFAR-100.

CIFAR-10							
	Near-OOD		Far-OOD		Avg.		
Method	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	Acc.↑
CE	40.2	88.8	30.6	91.0	35.4	89.9	**95.2**
NPOS	32.3	90.3	21.1	94.6	26.7	92.4	92.1
CIDER	31.4	90.1	22.5	93.4	26.9	91.7	92.8
LogitNorm	28.9	92.5	15.8	96.3	22.4	94.4	94.5
PALM	30.4	90.7	18.5	95.6	24.5	93.1	93.5
ReweightOOD	30.3	91.8	**10.2**	**97.6**	20.2	94.7	92.9
T2FNorm	25.8	93.0	15.2	96.5	20.5	94.8	94.7
HamOS	28.5	91.5	16.2	95.6	22.3	93.5	93.8
ASL (Ours)	**25.1**	**93.3**	12.0	97.2	**18.6**	**95.2**	94.8

CIFAR-100							
	Near-OOD		Far-OOD		Avg.		
Method	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	Acc.↑
CE	67.8	74.8	51.9	82.6	59.8	78.7	**77.4**
NPOS	67.1	77.7	50.2	83.6	58.6	80.7	72.9
CIDER	68.4	74.4	61.5	76.9	64.9	75.7	68.1
LogitNorm	62.4	78.5	50.4	82.1	56.4	80.3	76.1
PALM	64.7	77.2	**37.8**	86.7	**51.2**	82.0	75.7
ReweightOOD	65.8	76.5	42.0	**88.1**	53.9	82.3	71.5
T2FNorm	58.2	79.9	49.5	83.3	53.8	81.6	76.3
HamOS	62.7	79.1	54.2	80.6	58.5	79.9	74.9
ASL (Ours)	**56.2**	**80.9**	47.8	85.0	52.0	**83.0**	76.7

[53]. We also evaluate covariate shift via corrupted ID samples [17]. For safety-critical scenarios, we further conduct experiments on the BIMCV medical dataset [50], where near-OOD consists of RSNA Bone Age [13] and COVID-19 CT scans [56], while far-OOD samples are selected from mainstream natural image datasets, including MNIST, CIFAR-10, Texture, and TIN.

We employ a ResNet-18 [14] for experiments on CIFAR-10, CIFAR-100, ImageNet-200, and BIMCV. For ImageNet-1K, we utilize standard pre-trained ResNet-50 and ViT-B/16 [10] models from Torchvision [32]. Further detailed training configurations are deferred to Appendix Sec. 2.

Evaluation Metrics. We adopt two standard metrics for OOD detection: the Area Under the Receiver Operating Characteristic curve (AUROC) and the False Positive Rate at 95% True Positive Rate (FPR95). Higher AUROC and lower FPR95 values indicate better performance.

Baselines. We benchmark our method against a wide range of state-of-the-art approaches, which can be categorized as follows: 1) Post-hoc methods: MSP [18], MLS [16], MDS [25], RMDS [43], ReAct [45], VIM [53], ASH [9], KNN [46], fDBD [27], and NCI [28]. 2) Training-based methods: CIDER [34], LogitNorm [54], PALM [31], ReweightOOD [42], and T2FNorm [41].

Table 4: OOD detection performance under covariate shift on ImageNet-200 using a ResNet-18 model trained from scratch.

	Near-OOD						Far-OOD							
	SSB-hard		NINCO		Avg.		iNaturalist		Textures		OpenImage-O		Avg.	
Method	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑
CE	78.1	70.2	55.6	79.2	66.9	74.7	28.4	90.0	**22.1**	**95.1**	39.9	86.0	30.1	90.3
CIDER	75.2	68.0	53.0	78.0	64.1	73.0	29.1	88.9	26.0	94.0	36.4	85.2	30.5	89.4
LogitNorm	71.7	71.3	54.1	79.9	62.9	75.6	24.5	92.0	40.3	87.4	39.7	85.0	34.8	88.1
PALM	74.8	67.8	57.5	77.1	66.2	72.4	31.8	88.2	22.2	94.5	38.7	84.5	30.9	89.1
ReweightOOD	72.0	72.9	49.7	79.3	60.9	76.1	27.6	89.0	28.8	92.3	36.8	85.2	31.1	88.8
T2FNorm	71.5	72.8	54.2	80.1	62.9	76.5	**23.6**	92.4	43.8	83.8	39.1	84.7	35.5	86.9
ASL (Ours)	**70.7**	**74.9**	**47.2**	**82.3**	**59.0**	**78.6**	23.8	**92.5**	30.4	91.9	**33.1**	**88.0**	**29.1**	**90.8**

Table 5: OOD detection performance on ImageNet with fine-tuned ResNet-50 (Left), ViT-B/16 (Right).

ResNet-50							
	Near-OOD		Far-OOD		Avg.		
Method	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	Acc.↑
CE	73.1	71.4	26.6	93.9	49.8	82.7	76.2
CIDER	78.1	66.8	35.2	91.8	56.7	79.3	66.2
LogitNorm	69.5	74.0	30.8	91.8	50.2	82.9	76.8
PALM	68.8	72.4	21.0	**95.3**	**44.9**	83.9	70.3
ReweightOOD	70.7	71.3	**19.7**	95.2	45.2	83.3	60.5
T2FNorm	68.0	75.9	34.7	90.4	51.4	83.1	76.8
ASL (Ours)	**65.4**	**78.8**	28.3	93.3	46.9	**86.0**	**76.8**

ViT-B/16							
	Near-OOD		Far-OOD		Avg.		
Method	FP↓	AU↑	FP↓	AU↑	FP↓	AU↑	Acc.↑
CE	66.7	78.5	27.3	91.9	47.0	85.2	**81.1**
CIDER	71.9	76.1	37.2	90.7	54.6	83.4	77.6
LogitNorm	84.3	74.3	50.9	86.3	67.6	80.3	80.5
PALM	70.4	79.1	31.7	91.9	51.1	85.5	77.2
ReweightOOD	72.6	77.7	34.0	90.6	53.3	84.1	71.2
T2FNorm	83.2	76.1	73.7	83.4	78.5	79.8	80.5
ASL (Ours)	**63.3**	**81.9**	**21.2**	**94.0**	**42.2**	**88.0**	80.9

4.2 Main Result

ASL achieves state-of-the-art performance from scratch on ImageNet-200 and CIFAR benchmarks. We evaluate ASL under the standard setting of training from scratch on ImageNet-200 and CIFAR datasets. As shown in Tab. 2, ASL establishes a new state-of-the-art among training-based methods, outperforming the next-best method by 1.4 percentage points in average AUROC on near-OOD tasks. On CIFAR-10 and CIFAR-100 (Tab. 3), ASL consistently outperforms all competing methods, which shows ASL's effectiveness in learning a highly discriminative feature space without complex training objectives.

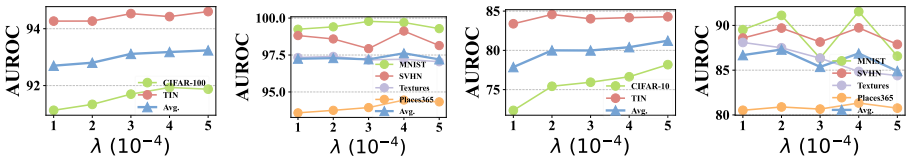
ASL sustains robust OOD detection performance under covariate shift. We further evaluate ASL under the covariate shift setting, where its robustness is particularly evident. As shown in Tab. 4, ASL demonstrates a significant advantage in detecting near-OOD samples on ImageNet-200. More strikingly, for far-OOD detection, ASL is the only training-based method that improves upon the standard cross-entropy baseline.

ASL significantly enhances OOD detection when fine-tuning pre-trained models on ImageNet. To assess ASL in a practical fine-tuning scenario, we apply it to pre-trained ResNet-50 and ViT-B/16 models on the full ImageNet-1K dataset. The results, presented in Tab. 5, show substantial improvements. For ResNet-50, fine-tuning with ASL significantly boosts the near-OOD AUROC from a baseline of 71.4% to 78.8%. For ViT-B/16, ASL

Table 6: OOD detection performance on BIMCV under standard (Left) and covariate shift settings (Right).

Method	Near-OOD FP↓	Near-OOD AU↑	Far-OOD FP↓	Far-OOD AU↑	Avg. FP↓	Avg. AU↑
CE	21.1	95.0	24.9	95.9	23.0	95.5
NPOS	63.3	72.8	56.9	72.3	60.1	72.6
CIDER	46.7	74.7	40.6	82.4	43.7	78.6
LogitNorm	97.2	33.7	98.5	37.8	97.9	35.8
PALM	86.2	49.6	88.9	47.6	87.6	48.6
ReweightOOD	38.7	90.9	13.3	97.1	26.0	94.0
T2FNorm	85.5	58.8	88.2	50.0	86.9	54.4
HamOS	88.1	52.2	80.2	62.2	84.2	57.2
ASL (Ours)	**15.8**	**96.3**	**8.99**	**98.1**	**12.4**	**97.2**

Method	Near-OOD FP↓	Near-OOD AU↑	Far-OOD FP↓	Far-OOD AU↑	Avg. FP↓	Avg. AU↑
CE	20.9	95.2	24.8	96.0	22.9	95.6
NPOS	63.3	72.8	56.3	72.4	59.8	72.6
CIDER	48.2	74.4	41.6	81.7	44.9	78.1
LogitNorm	97.5	33.0	98.6	37.1	98.1	35.1
PALM	86.1	49.5	89.3	47.6	87.7	48.6
ReweightOOD	39.3	91.0	13.2	97.2	26.3	94.1
T2FNorm	86.7	58.0	88.7	49.3	87.7	53.7
HamOS	88.0	52.5	80.2	62.5	84.1	57.5
ASL (Ours)	**15.2**	**96.4**	**8.58**	**98.2**	**11.9**	**97.3**



(a) CIFAR-10 (Near) **(b)** CIFAR-10 (Far) **(c)** CIFAR-100 (Near) **(d)** CIFAR-100 (Far)

Fig. 4: Ablation study on the weight decay hyperparameter λ. Performance of ASL across a wide range of λ values (1×10^{-4} to 5×10^{-4}) on CIFAR-10 and CIFAR-100.

concurrently improves near-OOD AUROC from 78.5% to 81.9% and far-OOD AUROC from 91.9% to a leading 94.0%.

ASL shows notable effectiveness and promising potential for safety-critical medical imaging. To validate its practical value in real-world scenarios, we evaluate ASL on the BIMCV dataset under both standard and covariate shift conditions (Tab. 6). In the standard setting, ASL achieves an outstanding average FPR95 of 12.4%, significantly outperforming the cross-entropy baseline's 23.0%. Crucially, under the covariate shift setting—which simulates realistic domain shifts across different hospitals—ASL maintains its robust performance. It reduces the near-OOD FPR95 to 15.2% (down from the baseline's 20.9%) and achieves a state-of-the-art average AUROC of 97.3%. These results highlight ASL's strong potential for safe deployment in domains where robust detection of unexpected anomalies is paramount.

Table 7: OOD Detection Performance of Post-hoc Methods with/without ASL on CIFAR-10.

Method	Near-OOD FP↓	Near-OOD AU↑	Far-OOD FP↓	Far-OOD AU↑
MSP	45.8	88.3	30.0	91.4
MLS	60.0	88.1	35.4	92.2
ReAct	62.0	86.7	42.0	91.1
ASH	89.2	73.8	76.9	77.9
MDS	49.2	85.2	33.3	89.0
+ASL	**25.1**	**93.3**	**12.0**	**97.2**
RMDS	38.6	89.8	25.9	92.4
+ASL	33.7	90.6	18.9	94.4
VIM	45.6	88.4	28.1	92.4
+ASL	26.2	93.0	13.1	96.9
KNN	34.4	90.7	24.4	93.1
+ASL	26.6	92.9	15.6	96.5
fDBD	35.5	90.6	24.3	93.4
+ASL	30.3	92.1	15.5	96.3
NCI	48.7	88.6	33.8	90.8
+ASL	34.1	91.3	14.7	96.4
MD++	40.2	88.8	30.6	91.0
+ASL	**25.1**	**93.3**	**12.0**	**97.2**

4.3 Analysis

Combine with Distance-Based Methods. As demonstrated in Tab. 7, applying various distance-based post-hoc methods to features extracted from an ASL-trained model consistently and significantly enhances their performance compared to using features from a standard cross-entropy model. For example, when paired with ASL features, VIM's near-OOD FPR95 plummets from 45.6% to 26.2%, and its far-OOD FPR95 is more than halved, decreasing from 28.1% to 13.1%.

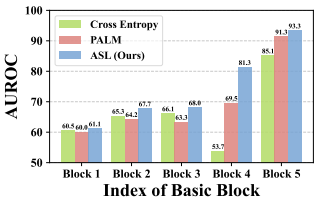


Fig. 5: OOD detection performance (AUROC) across different layers.

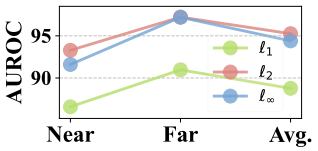


Fig. 6: Different Normalization.

Impact of Weight Decay Hyperparameter λ. As shown in Proposition 1, the weight decay hyperparameter λ constrains the norms of the classifier weights. We ablate λ in Fig. 4. Higher λ prompts the model to learn finer features, improving near-OOD detection, while far-OOD performance remains robust. Thus, λ effectively boosts near-OOD detection without harming far-OOD performance.

Effectiveness Progressively Increases Across Network Layers. To understand how ASL influences representation learning throughout the network, we evaluate OOD detection performance using features from different blocks of ResNet-18 on CIFAR-10. Fig. 5 clearly illustrates that ASL consistently learns more discriminative features than both standard cross-entropy and PALM at every network depth. Notably, in Block 4, ASL's performance shows a marked improvement to an AUROC of 81.3%. This demonstrates that ASL is already learning a highly effective and separable feature space before the final layer.

Table 8: Training time per epoch (in seconds) on ImageNet-200 for training-based methods.

Method	Time↓
CE	**252**$_{\pm 1.63}$
CIDER	657$_{\pm 4.55}$
NPOS	874$_{\pm 2.16}$
HamOS	3399$_{\pm 7.76}$
LogitNorm	294$_{\pm 1.25}$
PALM	381$_{\pm 2.36}$
ReweightOOD	406$_{\pm 2.36}$
T2FNorm	293$_{\pm 2.05}$
ASL (Ours)	<u>277</u>$_{\pm 1.70}$

Comparison of Norm-Based Normalization Schemes. We investigated the impact of different normalization schemes on out-of-distribution (OOD) detection performance, and compared ℓ_1, ℓ_2, and ℓ_∞ normalization on CIFAR-10. As shown in Fig. 6, ℓ_2 normalization achieves the best performance. This is consistent with our core motivation of optimizing features on the unit hypersphere. This geometric constraint forces the model to rely solely on angular information for discrimination, thereby reducing sensitivity to feature norms.

Training Efficiency. ASL offers significant practical advantages in training efficiency. As shown in Tab. 8, ASL maintains competitive training time per epoch (277s) on ImageNet-200, remaining close to the standard cross-entropy

baseline while being substantially faster than other training methods. This efficiency makes our approach not only effective but also highly practical for large-scale applications. Further experimental results and detailed discussions can be found in Appendix Sec. 5.

5 Conclusion

In this work, we propose ASL, a training framework that addresses a key challenge in OOD detection by applying ℓ_2-normalization to features before the cross-entropy loss. This simple yet powerful modification forces the model to learn optimal angular separation, significantly improving its ability to distinguish challenging near-OOD samples. Extensive experiments demonstrate that ASL achieves state-of-the-art performance on major benchmarks. It delivers these results with high training efficiency and proves effective in fine-tuning on large-scale datasets. We hope this work serves as a compelling baseline and encourages a re-evaluation of the complexity-performance Pareto frontier in future research.

Acknowledgements

This work was supported by the Ministry of Science and Technology of China under Grant No. 2025ZD0123800, the HUST Interdisciplinary Research Program under Grant No. 2025JCYJ077, and KingSoft 2026 University-Industry Project. We sincerely thank the anonymous reviewers for their constructive comments, particularly the detailed feedback on the ablation studies and theoretical justifications. This work has benefited substantially from their expertise.

References

1. Ammar, M.B., Belkhir, N., Popescu, S., Manzanera, A., Franchi, G.: NECO: NEural collapse based out-of-distribution detection. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=9ROuKblmi7
2. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016)
3. Bitterwolf, J., Mueller, M., Hein, M.: In or out? fixing imagenet out-of-distribution detection evaluation. arXiv preprint arXiv:2306.00826 (2023)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PmLR (2020)
5. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3606–3613 (2014)
6. Demirel, B., Fumero, M., Locatello, F.: OOD detection with relative angles. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=ul5mlXrLZb

7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. Deng, L.: The mnist database of handwritten digit images for machine learning research [best of the web]. IEEE signal processing magazine **29**(6), 141–142 (2012)
9. Djurisic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely simple activation shaping for out-of-distribution detection. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=ndYXTEL6cZz
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
11. Du, X., Wang, Z., Cai, M., Li, S.: Towards unknown-aware learning with virtual outlier synthesis. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=TW7d65uYu5M
12. Haas, J., Yolland, W., Rabus, B.T.: Linking neural collapse and l2 normalization with improved out-of-distribution detection in deep neural networks. Transactions on Machine Learning Research (2023), https://openreview.net/forum?id=fjkN5Ur2d6
13. Halabi, S.S., Prevedello, L.M., Kalpathy-Cramer, J., Mamonov, A.B., Bilbily, A., Cicero, M., Pan, I., Pereira, L.A., Sousa, R.T., Abdala, N., et al.: The rsna pediatric bone age machine learning challenge. Radiology **290**(2), 498–503 (2019)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. He, Z., Zhao, C., Shao, M., Wu, X., Zhao, X., Li, D., Tian, Q., Yu, L.: Out-of-distribution detection with positive and negative prompt supervision using large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 21699–21707 (2026)
16. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. In: International Conference on Machine Learning. pp. 8759–8773. PMLR (2022)
17. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019), https://openreview.net/forum?id=HJz6tiCqYm
18. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=Hkg4TI9xl
19. Henriksson, J., Ursing, S., Erdogan, M., Warg, F., Thorsén, A., Jaxing, J., Örsmark, O., Toftås, M.Ö.: Out-of-distribution detection as support for autonomous driving safety lifecycle. In: International Working Conference on Requirements Engineering: Foundation for Software Quality. pp. 233–242. Springer (2023)
20. Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. Advances in Neural Information Processing Systems **34**, 677–689 (2021)
21. Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M., Yi, X.: A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. Computer Science Review **37**, 100270 (2020)

22. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems **33**, 18661–18673 (2020)
23. Krizhevsky, A., et al.: Learning multiple layers of features from tiny images (2009)
24. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
25. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems **31** (2018)
26. Li, H., Zhang, T.: Outlier synthesis via hamiltonian monte carlo for out-of-distribution detection. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=N6ba2xsmds
27. Liu, L., Qin, Y.: Fast decision boundary based out-of-distribution detector. In: Proceedings of the 41st International Conference on Machine Learning. pp. 31728–31746 (2024)
28. Liu, L., Qin, Y.: Detecting out-of-distribution through the lens of neural collapse. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15424–15433 (2025)
29. Liu, W., Wang, X., Owens, J., Li, Y.: Energy-based out-of-distribution detection. Advances in neural information processing systems **33**, 21464–21475 (2020)
30. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)
31. Lu, H., Gong, D., Wang, S., Xue, J., Yao, L., Moore, K.: Learning with mixture of prototypes for out-of-distribution detection. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=uNkKaD3MCs
32. maintainers, T., contributors: Torchvision: Pytorch's computer vision library. https://github.com/pytorch/vision (2016)
33. Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., Li, Y.: Delving into out-of-distribution detection with vision-language representations. Advances in neural information processing systems **35**, 35087–35102 (2022)
34. Ming, Y., Sun, Y., Dia, O., Li, Y.: How to exploit hyperspherical embeddings for out-of-distribution detection? In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=aEFaE0W5pAd
35. Müller, M., Hein, M.: Mahalanobis++: Improving OOD detection via feature normalization. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=vutMcZl50l
36. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y., et al.: Reading digits in natural images with unsupervised feature learning. In: NIPS workshop on deep learning and unsupervised feature learning. vol. 2011, p. 7. Granada (2011)
37. Park, J., Chai, J.C.L., Yoon, J., Teoh, A.B.J.: Understanding the feature norm for out-of-distribution detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1557–1567 (2023)
38. Park, J., Jung, Y.G., Teoh, A.B.J.: Nearest neighbor guidance for out-of-distribution detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1686–1695 (2023)
39. Rabe, M., Milz, S., Mader, P.: Development methodologies for safety critical machine learning applications in the automotive domain: A survey. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 129–141 (2021)

40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
41. Regmi, S., Panthi, B., Dotel, S., Gyawali, P.K., Stoyanov, D., Bhattarai, B.: T2fnorm: Train-time feature normalization for ood detection in image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 153–162 (2024)
42. Regmi, S., Panthi, B., Ming, Y., Gyawali, P.K., Stoyanov, D., Bhattarai, B.: Reweightood: Loss reweighting for distance-based ood detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 131–141 (2024)
43. Ren, J., Fort, S., Liu, J., Roy, A.G., Padhy, S., Lakshminarayanan, B.: A simple fix to mahalanobis distance for improving near-ood detection. arXiv preprint arXiv:2106.09022 (2021)
44. Sehwag, V., Chiang, M., Mittal, P.: SSD: A unified framework for self-supervised outlier detection. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=v5gjXpmR8J
45. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. Advances in neural information processing systems **34**, 144–157 (2021)
46. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution detection with deep nearest neighbors. In: International conference on machine learning. pp. 20827–20840. PMLR (2022)
47. Tao, L., Du, X., Zhu, J., Li, Y.: Non-parametric outlier synthesis. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=JHklpEZqduQ
48. Techapanurak, E., Suganuma, M., Okatani, T.: Hyperparameter-free out-of-distribution detection using cosine similarity. In: Proceedings of the Asian conference on computer vision (2020)
49. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
50. Vayá, M.D.L.I., Saborit, J.M., Montell, J.A., Pertusa, A., Bustos, A., Cazorla, M., Galant, J., Barber, X., Orozco-Beltrán, D., García-García, F., et al.: Bimcv covid-19+: A large annotated dataset of rx and ct images from covid-19 patients. arXiv preprint arXiv:2006.01174 (2020)
51. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-set recognition: A good closed-set classifier is all you need. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=5hLP5JY9S2d
52. Wang, F., Xiang, X., Cheng, J., Yuille, A.L.: Normface: L2 hypersphere embedding for face verification. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1041–1049 (2017)
53. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4921–4930 (2022)
54. Wei, H., Xie, R., Cheng, H., Feng, L., An, B., Li, Y.: Mitigating neural network overconfidence with logit normalization. In: International conference on machine learning. pp. 23631–23644. PMLR (2022)

55. Winkens, J., Bunel, R., Roy, A.G., Stanforth, R., Natarajan, V., Ledsam, J.R., MacWilliams, P., Kohli, P., Karthikesalingam, A., Kohl, S., et al.: Contrastive training for improved out-of-distribution detection. arXiv preprint arXiv:2007.05566 (2020)
56. Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., Xie, P.: Covid-ct-dataset: a ct scan dataset about covid-19. arXiv preprint arXiv:2003.13865 (2020)
57. Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Li, Y., Liu, Z., Chen, Y., Li, H.: OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. Journal of Data-centric Machine Learning Research (2024), https://openreview.net/forum?id=cnnTnJQigs, dataset Certification
58. Zhang, Z., Xiang, X.: Decoupling maxlogit for out-of-distribution detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3388–3397 (2023)
59. Zhang, Z., Xu, Z., Xiang, X.: Vision-language dual-pattern matching for out-of-distribution detection. In: European Conference on Computer Vision. pp. 273–291. Springer (2024)
60. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)
61. Zimmerer, D., Full, P.M., Isensee, F., Jäger, P., Adler, T., Petersen, J., Köhler, G., Ross, T., Reinke, A., Kascenas, A., et al.: Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. IEEE transactions on medical imaging **41**(10), 2728–2738 (2022)