# Spurious-Aware Prototype Refinement for Reliable Out-of-Distribution Detection

**Reihaneh Zohrabi**[*1] **Hosein Hasani**[*2] **Mahdieh Soleymani Baghshah**[2] **Anna Rohrbach**[1]
**Marcus Rohrbach**[1] **Mohammad Hossein Rohban**[2]
[1]TU Darmstadt [2]Sharif University of Technology
{reihaneh.zohrabi, anna.rohrbach, marcus.rohrbach}@tu-darmstadt.de
{hosein.hasani, soleymani, rohban}@sharif.edu

## Abstract

Out-of-distribution (OOD) detection is crucial for ensuring the reliability and safety of machine learning models in real-world applications, where they frequently face data distributions unseen during training. Despite progress, existing methods are often vulnerable to spurious correlations that mislead models and compromise robustness. To address this, we propose SPROD, a novel prototype-based OOD detection approach that explicitly addresses the challenge posed by unknown spurious correlations. Our post-hoc method refines class prototypes to mitigate bias from spurious features without additional data or hyperparameter tuning, and is broadly applicable across diverse backbones and OOD detection settings. We conduct a comprehensive spurious correlation OOD detection benchmarking, comparing our method against existing approaches and demonstrating its superior performance across challenging OOD datasets, such as CelebA, Waterbirds, UrbanCars, Spurious Imagenet, and the newly introduced Animals MetaCoCo. On average, SPROD improves AUROC by 4.8% and FPR@95 by 9.4% over the second best.

## 1 Introduction

Machine learning systems in real-world applications often encounter out-of-distribution (OOD) inputs, which are samples from distributions different from the training data. These inputs require cautious handling to prevent overconfident mispredictions during inference [1]. This makes OOD detection crucial, as it aims to identify whether an input belongs to the known distribution or not. Yet, deep neural networks, widely used in vision tasks [2–5], tend to make high-confidence predictions even on OOD inputs, demonstrating their inability to recognize data outside the training distribution as OOD [6, 7]. The reliability of OOD detection is especially critical in applications like healthcare and autonomous driving, where overconfident predictions on unfamiliar data could have serious consequences [8, 9].

Recent research on OOD detection aims to ensure the reliable deployment of DNNs [7, 10–13]. Despite many effective methods, their robustness can be undermined by *spurious correlations in the training data* [14]. Studies indicate that models often rely on features that are statistically informative but not causally representative of the object itself [15–17]. These misleading cues can act as shortcuts, allowing models to achieve high accuracy without learning the core, causally relevant features [18]. While spurious correlations have been well-explored in classification tasks [17, 19, 20], their impact on OOD detection remains underexplored. Recently, [14] underscores the impact of spurious correlations on OOD detection and introduces a formalization that categorizes OOD samples into two types: spurious OODs (SP-OODs), which contain spurious attributes but lack core features,

[*]Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

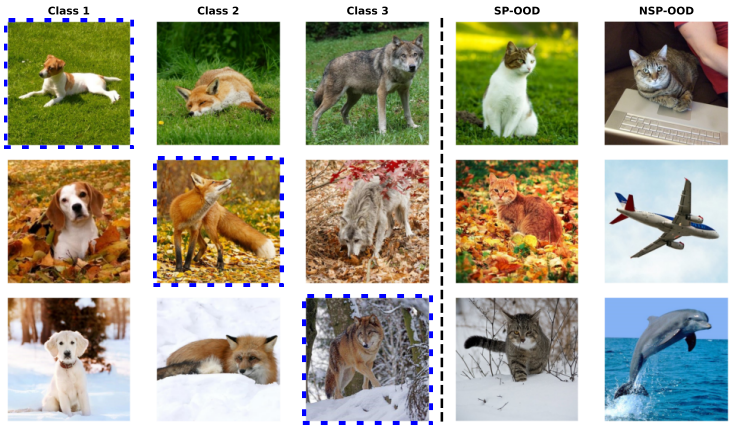



Figure 1: The challenge of spurious correlations in OOD detection. ID classes (dog, fox, wolf) appear in correlated backgrounds (grass, autumn, snow), with majority groups relying on context shortcuts (blue frames). SP-OOD samples share the same contextual backgrounds, making detection more difficult. NSP-OOD samples differ in context and lack both spurious and core features.

and non-spurious OODs (NSP-OODs), which lack both attributes and align with the traditional OOD setting. Figure 1 shows an example of this problem setting.

Recent efforts to mitigate spurious correlations in OOD detection can be grouped into several broad categories. Outlier Exposure (OE) techniques reduce reliance on spurious correlations by incorporating synthetic OOD samples in their training [21–24]. Other methods focus on modifying training objectives to explicitly discourage models from depending on spurious features [25, 26]. However, these methods usually require retraining and generating additional OOD data. In contrast, several *post-hoc* approaches have been proposed that bypass the limitations of training-heavy approaches and offer fast and light alternatives [27–32]. Despite their promise, these methods are often tested on limited synthetic datasets or specific backbones, and some rely on multiple modalities.

To address these limitations, we propose **S**purious-Aware **P**rototype **R**efinement for Reliable **O**ut-of-Distribution **D**etection (SPROD) for robust OOD detection, especially in the presence of unknown spurious correlations. It follows a three-stage process: (1) initial prototype construction, (2) classification-aware prototype calculation, and (3) group prototype refinement. SPROD can be easily applied to any pretrained feature extractor without fine-tuning on target datasets, offering a straightforward, hyperparameter-free approach that is both efficient and adaptable across diverse OOD detection tasks. Moreover, our work offers a comprehensive evaluation of OOD detection in the presence of spurious correlations, benchmarking existing methods across multiple challenging datasets, including Waterbirds [33], CelebA [34], UrbanCars [35], Spurious ImageNet [36], and the newly introduced Animals MetaCoCo. SPROD achieves state-of-the-art performance and consistently exhibits robust behavior across a wide range of benchmarks and experimental conditions. The main contributions of this work are as follows:

- We propose SPROD, a post-hoc OOD detection method that directly addresses unknown spurious correlations by design and outperforms the state-of-the-art. We provide theoretical insight into how the proposed method mitigates spurious bias.
- SPROD is a fast, simple, and general approach applicable to diverse pretrained feature extractors and OOD detection settings.
- SPROD does not assume access to group annotations to achieve robustness to spurious correlations and does not require either OOD or ID validation data for hyperparameter tuning. Moreover, it maintains strong performance even in low-training data regimes.
- This work conducts and introduces comprehensive benchmarking across multiple SP-OOD datasets, including the newly introduced Animals MetaCoCo, a realistic, multiclass dataset with diverse spurious attributes.
- Finally, our study sheds new light on key factors influencing SP-OOD detection, such as the impact of backbone fine-tuning and the choice of scoring mechanisms.

## 2 Related Work

OOD detection methods can be categorized into training-time and post-hoc approaches [37]. Training-time methods leverage auxiliary OOD samples (Outlier Exposure) [12, 38, 39] or apply regularization [40–43] to enhance OOD detection. Post-hoc methods, in contrast, derive OOD scores from base classifiers without modifying training [37]. Overall, post-hoc methods offer simplicity and competitive performance [37], making them practical under limited data or training resources. Among post-hoc methods, several approaches apply transformations to model logits to derive OOD scores. MSP [7] uses the maximum softmax probability, the energy-based method [13] computes the log-sum-exp of logits, MLS [44] uses the maximum logit and introduces KL Matching (KLM) based on KL divergence, and GEN [45] employs generalized entropy of softmax outputs.

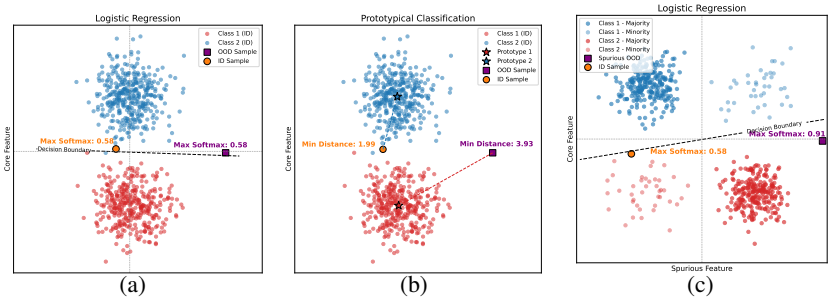
Another class of post-hoc methods detects OOD samples via feature-space distances. MDS [11] fits class-conditional Gaussians to pre-logit features and computes Mahalanobis distances, refined by RMDS [46] with an unconditional Gaussian on ID data. KNN [47] uses distances to nearest ID samples. SHE [48] scores samples by their distance to stored ID feature templates. NNGuide [49] leverages nearest-neighbor guidance to adjust test features toward the ID manifold. Relation [50] constructs a graph over training embeddings and detects outliers via relational anomalies. NECO [51] scores samples by their feature alignment with class weight vectors, leveraging neural collapse geometry. SCALE [52] separates ID and OOD samples by scaling penultimate-layer activations. FDBD [53] measures features' regularized mean distance to the classifier's decision boundaries. NCI [54] scores samples by their distance to class weight vectors, filtered by feature norms.

Prototype-based methods shape class representations for distance-based OOD scoring. Classical approaches like MDS [11] and its variants [46] are closely related, as they model each class by a centroid in feature space, optionally using class-conditional covariances to compute Mahalanobis distances. Recent works extend this via explicit training objectives. CIDER [55] learns hyperspherical embeddings by jointly enforcing intra-class compactness and inter-class dispersion, thereby improving ID and OOD separability. PALM [56] represents each class as a mixture of learnable prototypes and optimizes a maximum-likelihood and contrastive objective, updating prototypes and backbone features jointly during training. PROWL [57] also leverages prototype representations, but for pixel-level OOD detection in segmentation. While these methods share a prototypical framework, *SPROD* differs as a post-hoc method operating on pretrained backbones and is explicitly designed to mitigate the negative effects of unknown spurious correlations. Appendix H further analyzes a variant, SPROD-KMeans, which connects to mixture-of-prototypes ideas while remaining fully post-hoc.

A few methods combine information from both feature and logit spaces. ReAct [58] thresholds activations before applying energy-based scoring. ViM [59] adds a virtual logit from the residual norm between input features and the ID subspace and applies softmax over extended logits. ASH [60] prunes high-magnitude activations and rescales remaining features before logit computation, improving energy-based OOD separability. Some methods also exploit gradient space for OOD scoring [61, 62]. GradNorm [61] computes the KL divergence to a uniform distribution and uses the gradient norm (w.r.t. the penultimate layer) as the score.

Spurious correlations in training data degrade OOD detection performance, as shown in [14]. Evaluations of popular methods [7, 10, 11, 13, 63] reveal that as spurious correlations increase, detection performance drops, and SP-OOD samples become especially challenging to detect. Feature-based methods like MDS [11] outperform others, especially for NSP samples. Recent work addresses spurious correlations in OOD detection through various strategies. OE methods synthesize OOD samples in ways that reduce reliance on background cues, encouraging models to focus on core semantic features [21–24]. Training-time regularization mitigates spurious cues by reweighting samples or augmenting non-semantic features [25, 26]. Post-hoc methods improve inference by modifying inputs to isolate semantics or reduce background influence [27, 28].

Recent advances in vision-language models [64] have led to a category of zero-shot OOD detection methods [29–32] that use textual inputs, such as class names or attribute descriptions, to define ID data and identify OOD samples. While some report results on Waterbirds SP-OOD, this is not their main focus. Moreover, lacking training data with spurious correlations, their zero-shot setting does not fully capture the SP-OOD challenge. In contrast to these approaches, which rely on explicit text for OOD scoring, our method refines visual prototypes using only ID features and class labels from the training set. A detailed review of studies regarding SP-OOD can be found in Appendix B.




(a) (b) (c)

Figure 2: **(a)** A far-OOD sample may receive a high softmax score, similar to a near-boundary ID sample. **(b)** Distances to class prototypes offer a more consistent separation of OOD samples. **(c)** In the SP-OOD setting, the problem is even more severe: A biased decision boundary causes the OOD sample to receive high softmax confidence, while a minority ID sample receives lower confidence.

# 3 Preliminaries

## 3.1 Problem Setup

This paper addresses Out-of-Distribution (OOD) detection under spurious correlations in a supervised classification setting. Let $\mathcal{X}$ be the input space and $\mathcal{Y} = \{1, \ldots, C\}$ the label set. The in-distribution (ID) training data $\mathcal{D}_{\text{in}} = \{(x_i, y_i)\}_{i=1}^{N}$ comprises samples from a joint distribution $P_{\mathcal{X},\mathcal{Y}}$. A neural network $f_\theta$ maps each input $x_i$ to a feature embedding $h_i = f_\theta(x_i)$. This network is typically pretrained or fine-tuned on the training data. OOD detection aims to identify test samples from distributions not seen during training, including those from unseen classes.

Spurious correlations in OOD detection were first formalized by [14]. According to this framework, each input can be decomposed into: **(i)** *core features*, which are causally related to the label, and **(ii)** *spurious features*, which are correlated with the label but not causally relevant. Imbalances in the training data often result in dominant core–spurious combinations (*majority groups*), while rarer combinations (*minority groups*) remain underrepresented. This bias encourages models to rely disproportionately on spurious cues. In controllable settings, the proportion of majority group samples within a class is captured by the **correlation rate**. As illustrated in Figure 1, this setup gives rise to two types of OOD examples: **Spurious OOD (SP-OOD)**, which share spurious features with ID data but differ in core features (e.g., a *cat* on *grass*, where *grass* is spuriously associated with ID classes *dog, wolf, fox*); and **Non-spurious OOD (NSP-OOD)**, which differ in both core and spurious features (e.g., a *cat* on a *laptop*).

## 3.2 Score Calculation

A key element in OOD detection is the scoring function $\mathcal{S}(x)$, which assigns a scalar value reflecting how likely an input belongs to the in-distribution (ID). This score is typically derived from a model's learned representations or its predictive outputs. OOD detection is performed based on this score function. An effective OOD method offers distinct and well-separated distributions of scores for ID and OOD samples.

Prior studies show that feature-based scores, derived from intermediate representations [10, 13, 47], typically outperform those based on output probabilities [7], especially in the presence of spurious correlations [14]. Our experiments further support this trend in certain settings, suggesting that output-based methods (especially those relying on softmax probabilities) may be less reliable when model confidence is influenced by spurious features. To better understand this difference, consider two probabilistic perspectives in classification: discriminative models directly estimate $p(y|x)$ and focus on separating classes, while generative models estimate $p(x|y)$ and capture how data $x$ is distributed for each class $y$. Most distance-based OOD detection methods can be viewed as approximating a generative approach in feature space.

While directly modeling $p(y|x)$ is generally effective for classifying ID samples in standard settings, discriminative approaches have been reported to be more sensitive under distribution shifts, such as in continual learning [65] or in the presence of spurious correlations [66]. Furthermore, their utility for

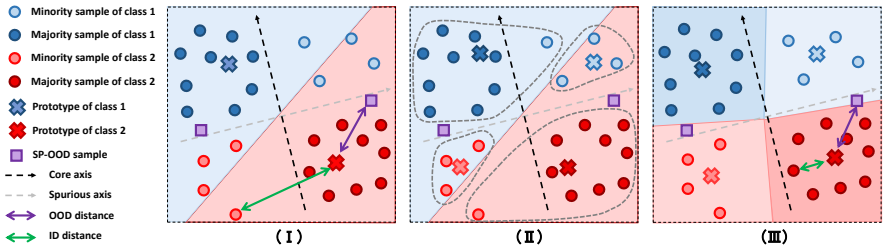



Figure 3: Overview of the three main stages of SPROD. In the first stage, class prototypes are computed, though they may be biased due to spurious correlations. In the second stage, group prototypes are constructed for the misclassified and correctly classified samples of each class. Finally, in the third stage, class samples are reassigned to their nearest group prototypes, and based on these assignments, refined minority and majority prototypes are recalculated.

OOD detection can be limited, particularly when methods are optimized solely on ID data without exposure to OOD examples. The softmax function forces a normalized probability distribution over the known classes, which can degrade OOD detection performance. For instance, an OOD sample that lies far from all class distributions may receive a high softmax probability if it is only slightly distant from a decision boundary. Conversely, an ID sample situated near the boundary between multiple classes could receive a low maximum softmax probability, potentially leading to its misclassification as OOD. In contrast, distance-based approaches, which rely on the class-conditional distribution $p(x|y)$, are by nature more robust in these scenarios. OOD samples that share few characteristics with any known class typically exhibit low likelihood under all class-conditional distributions $p(x|y)$, and can be reliably identified, regardless of their proximity to decision boundaries.

Figure 2 highlights the limitations of softmax-based OOD scoring in a controlled toy dataset. The challenge becomes more pronounced in the presence of spurious correlations, as shown in Figure 2c. In this scenario, a discriminative model, potentially biased by spurious features, may assign high confidence to an SP-OOD sample that shares these spurious cues with an ID class, while simultaneously assigning low confidence to a minority ID sample that lacks them. Motivated by these observations, we design a generative distance-based approach that is more robust to unknown spurious correlations.

# 4 SPROD

In this section, we introduce Spurious-Aware Prototype Refinement for Reliable Out-of-Distribution Detection (SPROD). SPROD adapts the prototypical framework [67] for robust OOD detection by constructing class prototypes designed to be resilient to spurious correlations. The core method involves a three-stage process, which is shown in Figure 3 and detailed in the following subsections.

## 4.1 Stage 1: Initial Prototype Construction

Given a pretrained feature extractor $f_\theta$, we first obtain feature embeddings $h_i = f_\theta(x_i)$ for each training sample $x_i$. To ensure uniformity in feature representation, these embeddings are normalized to have unit norm $z_i = h_i/\|h_i\|_2$. For each ID class $c \in \mathcal{Y}$, an initial prototype $\mathbf{p}_c$ is computed as the mean of these normalized embeddings: $\mathbf{p}_c = 1/N_c \sum_{i:y_i=c} z_i$, where $N_c$ is the number of samples in class $c$. Each $\mathbf{p}_c$ serves as an initial estimate of the class centroid in the normalized feature space. A query sample $x_q$ (with normalized embedding $z_q$) is typically classified to the class $c$ whose prototype $\mathbf{p}_c$ is the closest (e.g., using Euclidean distance $d(z_q, \mathbf{p}_c)$). We use $d(z_q, \mathbf{p}_c)$ as the scoring function for OOD detection. While our empirical results demonstrate the effectiveness of the naïve prototypical method in both SP-OOD and NSP-OOD settings, this approach remains vulnerable to biases from spurious correlations in the training data. As a result, the prototypes become skewed toward majority groups within each class (see part I of Figure 3). This bias leads to a scenario where SP-OOD samples (represented by purple squares in Figure 3) may be erroneously classified as ID due to their proximity to these biased prototypes. These limitations motivate the subsequent refinement stages.

### 4.2 Stage 2: Classification-Aware Prototype Calculation

To mitigate biases from spurious correlations present in the Stage 1 prototypes, we begin by analyzing how these initial prototypes classify the training data itself. Inspired by [20] , our debiasing process starts with classifying the training samples based on the initial prototypes and partitioning the samples based on their prediction outcomes. For each class $c$, this identifies a set of correctly classified samples, $\mathcal{S}_c^{\text{corr}}$, and (multiple) sets of misclassified samples, $\{\mathcal{S}_{c\to m}^{\text{misc}}\}_{m\neq c}$, where $m$ is the incorrectly predicted class. The core assumption is that samples in $\mathcal{S}_{c\to m}^{\text{misc}}$ belong to subgroups of class $c$ that share spurious features with class $m$, leading to their misclassification.

Formally, for each class $c$, we compute the prototype $\mathbf{p}_c^{\text{corr}}$ by averaging over embeddings of correctly classified samples $\mathcal{S}_c^{\text{corr}}$. For the misclassified samples of class $c$, we compute the set of misclassified group prototypes $\{\mathbf{p}_{c\to m}^{\text{misc}}\}_{m\neq c}$ by averaging over samples in $\{\mathcal{S}_{c\to m}^{\text{misc}}\}_{m\neq c}$ individually. The number of misclassified group prototypes for class $c$, denoted by $C_c^{\text{misc}}$, corresponds to the number of other classes that training samples from class $c$ have been misclassified as during evaluation.

This procedure expands the number of prototypes per class from 1 up to $C$ (total number of classes), helping incorporate diverse subgroup characteristics within each class. However, this approach has a potential limitation. Specifically, samples within the minority group may still contribute to the prototype for the correctly classified (majority) group if they were initially classified correctly, resulting in slightly biased prototypes (see part II of Figure 3). Hence, we further refine the group prototypes in the third stage.

### 4.3 Stage 3: Group Prototype Refinement

In the third stage, we refine the group prototypes computed in Stage 2 to further reduce the remaining bias within them. Inspired by the reassignment step in K-means clustering, we first reassign samples within each class $\{z_i \mid y_i = c\}$ to either majority $\mathcal{S}_c^{\text{maj}}$ or minority groups $\{\mathcal{S}_{c\to m}^{\text{min}}\}_{m\neq c}$ based on their proximity to the corresponding prototypes ( $\mathbf{p}_c^{\text{corr}}$ for majority members and $\{\mathbf{p}_{c\to m}^{\text{misc}}\}_{m\neq c}$ for minority members). Following this reassignment, refined prototypes are computed as the mean of the updated group members:

$$\mathbf{p}_c^{\text{maj}} = \frac{1}{|\mathcal{S}_c^{\text{maj}}|}\sum_{z_i\in\mathcal{S}_c^{\text{maj}}} z_i, \quad \mathbf{p}_{c\to m}^{\text{min}} = \frac{1}{|\mathcal{S}_{c\to m}^{\text{min}}|}\sum_{z_i\in\mathcal{S}_{c\to m}^{\text{min}}} z_i \quad \forall m \neq c \tag{1}$$

The refined prototypes $\mathbf{p}_c^{\text{maj}}$ and $\mathbf{p}_{c\to m}^{\text{min}}$, further reduce bias through the proposed refitting process. During classification and OOD detection, the query embedding $z_q$ is compared to all group-specific prototypes, and the final prediction is based on the nearest prototype, regardless of group type. This multiple-prototype approach reduces the likelihood of OOD samples being erroneously classified due to shared spurious attributes with any single prototype (see part III of Figure 3). For each sample, the OOD score is simply calculated based on the distance to the nearest group prototype.

**Algorithm 1** Spurious-Aware Prototype Refinement

**Input:** Training samples $\{(x_i, y_i)\}_{i=1}^N$, feature extractor $f_\theta$
**Output:** Refined class prototypes
Get feature embedding $h_i = f_\theta(x_i)$ and $z_i = \frac{h_i}{\|h_i\|_2} \forall x_i$
**for** each class $c = 1, \ldots, C$: ▷ **Stage 1: Constructing class prototypes**
    $\mathbf{p}_c = \frac{1}{N_c}\sum_{i:y_i=c} z_i$
Classify all training samples using initial prototypes
**for** each class $c = 1, \ldots, C$: ▷ **Stage 2: Augmenting class prototypes**
    Separate samples into correctly classified $\mathcal{S}_c^{\text{corr}}$ and misclassified $\{\mathcal{S}_{c\to m}^{\text{misc}}\}_{m\neq c}$
    Compute $\mathbf{p}_c^{\text{corr}}$ based on $\mathcal{S}_c^{\text{corr}}$ and $\{\mathbf{p}_{c\to m}^{\text{misc}}\}_{m\neq c}$ based on $\{\mathcal{S}_{c\to m}^{\text{misc}}\}_{m\neq c}$
**for** each class $c = 1, \ldots, C$: ▷ **Stage 3: Refining group prototypes**
    Construct majority $\mathcal{S}_c^{\text{maj}}$ and minority $\{\mathcal{S}_{c\to m}^{\text{min}}\}_{m\neq c}$ groups based on proximity to $\mathbf{p}_c^{\text{corr}}$ or $\{\mathbf{p}_{c\to m}^{\text{misc}}\}_{m\neq c}$
    Compute $\mathbf{p}_c^{\text{maj}}$ based on $\mathcal{S}_c^{\text{maj}}$ and $\{\mathbf{p}_{c\to m}^{\text{min}}\}_{m\neq c}$ based on $\{\mathcal{S}_{c\to m}^{\text{min}}\}_{m\neq c}$

The overall procedure of our prototype refinement strategy is outlined in Algorithm 1 and a theoretical justification of how the proposed procedure mitigates spurious bias is provided in Appendix A.

# 5 Experiments

## 5.1 Experimental Setup

We evaluate our method against a comprehensive suite of 19 post-hoc OOD detection approaches: MSP [7], MDS [11], RMDS [46], Energy [13], GradNorm [61], ReAct [58], MaxLogit [44], MLS & KLM [44], VIM [59], KNN [47], SHE [48], ASH [60], NECO [51], NNGuide [49], Relation [50], SCALE [52], fDBD [53], and NCI [54]. We assess performance primarily using the Area Under the Receiver Operating Characteristic curve (AUROC), a threshold-independent metric, and the False Positive Rate at 95% True Positive Rate (FPR@95). Additional metrics, including the Area Under the Precision-Recall curve (AUPR), are provided in Appendix G. We repeat all experiments five times with different random seeds and report the mean and standard deviation.

In this experiments section, we focus on the more challenging SP-OOD scenarios, particularly those with the highest degree of spurious correlation in each dataset, as our method achieves near-perfect performance on far NSP-OOD samples. Detailed results for the NSP-OOD setting are provided in Appendix E. Further analyses of SPROD's stages, along with broader evaluations across various transformer-based and convolutional backbones, are included in Appendix F and Appendix G. For consistency, the results in this section primarily use the widely adopted ResNet-50 [68] backbone, with ResNet-18 additionally used in one analysis. Beyond SP-OOD benchmarks, we also evaluate our approach under conventional (standard) OOD settings to further demonstrate its general applicability.

## 5.2 Datasets

Table 1: Overview of datasets used for SP-OOD evaluation. "# Groups" denotes the number of distinct subpopulations based on class and spurious attribute combinations. "NA" indicates cases where such grouping is not explicitly defined.

| Dataset | Type | # Classes | # Spurious Attr. | # Groups | SP-OOD |
|---|---|---|---|---|---|
| Waterbirds (WB) [33] | Synthetic | 2 (Bird Type) | 2 Backgrounds | 4 | Places [69] background |
| CelebA (CA) [34] | Real-world | 2 (Hair Color) | 2 Genders | 4 | Bald male (no hair) |
| UrbanCars (UC) [35] | Synthetic | 2 (Car Type) | 2 Backgrounds $\times$ 2 Objects | 8 | Background / Background + Object |
| Animals MetaCoCo (AMC) [ours] | Real-world | 24 (Animal Type) | 8 Backgrounds | NA | Leave-2-out (class-based) |
| Sp-ImageNet100 (SpI) [36] | Real-world | 100 (ImageNet classes) | NA (spurious visual features) | NA | Spurious ImageNet [36] |

For evaluating SP-OOD detection, we utilize five diverse datasets, whose properties are summarized in Table 1. Additional details and visual examples for all datasets, including the NSP datasets and their evaluation setup, are provided in Appendix D. The datasets include **Waterbirds (WB)** [33] and **CelebA (CA)** [34]. While widely adopted, these datasets present limitations in terms of scale and realism (CelebA, in particular, is noted for its label noise [70]), making them insufficient for comprehensive evaluation. To address this, our evaluation incorporates three additional datasets designed to test robustness under diverse conditions, including multi-class scenarios, multiple spurious attributes, and real-world complexity.

**UrbanCars (UC)** [35] is a binary classification dataset (urban vs. country cars) with two spurious attributes: background and a co-occurring object, both correlated with the class, making it a challenging multi-spurious benchmark. Next, we introduce **Animals MetaCoCo (AMC)**, a realistic SP-OOD benchmark created by subsampling and curating animal categories from MetaCoCo [71]. It contains 26 distinct classes, each with 8 different background types serving as imbalance shortcut attributes. We use a class-level leave-2-out setting, where two classes are held out as SP-OOD and the rest are treated as ID. The significant similarity in background distributions between the ID and OOD splits makes Animals MetaCoCo a particularly challenging multi-class benchmark for SP-OOD detection (Figure 1). Last, we also use the **Spurious ImageNet** dataset introduced in [36] as SP-OOD data. This dataset consists of real-world images that contain only spurious features, such as bird feeders or graffiti, without the actual class objects, yet are consistently misclassified as specific ImageNet classes. These OOD samples are constructed for a subset of 100 ImageNet classes identified to rely on harmful spurious correlations. We refer to this subset of classes as Sp-ImageNet100 (SpI), a name we use for clarity, and treat it as the ID data.

For conventional OOD evaluations, we follow the same settings as in OpenOOD [37]. Specifically, we use **CIFAR-10** [72], **CIFAR-100** [72], and **ImageNet-1k** [73], along with their respective near-OOD and far-OOD datasets, to ensure consistency with widely adopted benchmarks in the field.

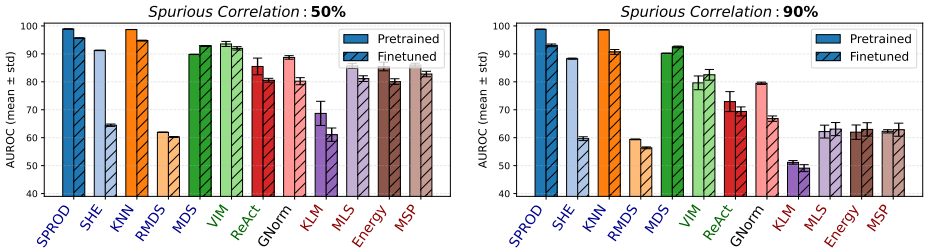
Table 2: Comparative performance of post-hoc OOD detection methods on SP-OOD benchmarks using a ResNet-50 backbone. Left: AUROC scores (higher is better); Right: FPR@95 scores (lower is better). Feature-based methods are indicated in blue, output-based methods in red, gradient-based in black, and hybrid methods in green. For each experiment, the top-performing method is shown in **bold**, and the second-best is underlined.

**AUROC↑**

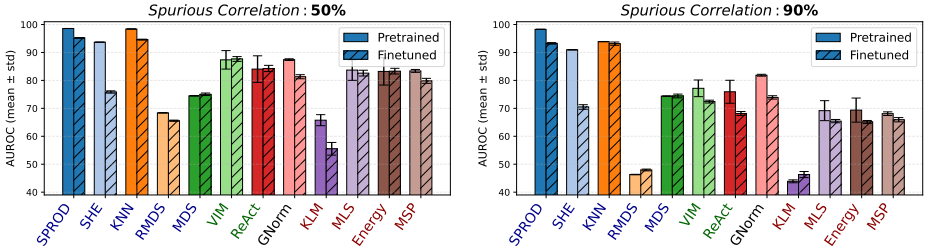
| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP[7] | $62.3_{\pm 0.6}$ | $46.0_{\pm 1.4}$ | $38.5_{\pm 0.3}$ | $79.7_{\pm 0.4}$ | $83.1_{\pm 0.3}$ | 61.9 |
| Energy[13] | $62.0_{\pm 2.6}$ | $45.4_{\pm 3.4}$ | $38.4_{\pm 2.1}$ | $79.9_{\pm 0.6}$ | $80.6_{\pm 0.4}$ | 61.3 |
| MLS[44] | $62.2_{\pm 2.3}$ | $45.3_{\pm 3.2}$ | $38.4_{\pm 1.4}$ | $80.2_{\pm 0.6}$ | $81.9_{\pm 0.3}$ | 61.6 |
| KLM[44] | $51.2_{\pm 0.7}$ | $41.7_{\pm 2.5}$ | $57.0_{\pm 0.2}$ | $74.2_{\pm 0.6}$ | $79.6_{\pm 0.8}$ | 60.7 |
| GEN[45] | $62.3_{\pm 0.6}$ | $46.0_{\pm 1.4}$ | $38.5_{\pm 0.3}$ | $80.2_{\pm 0.0}$ | $80.8_{\pm 0.4}$ | 61.6 |
| GNorm[61] | $79.5_{\pm 0.4}$ | $38.0_{\pm 1.3}$ | $46.6_{\pm 0.4}$ | $74.2_{\pm 0.5}$ | $85.2_{\pm 0.2}$ | 64.7 |
| ReAct[58] | $72.9_{\pm 3.6}$ | $45.6_{\pm 5.3}$ | $41.3_{\pm 3.1}$ | $80.1_{\pm 0.6}$ | $83.6_{\pm 0.7}$ | 64.7 |
| VIM[59] | $79.6_{\pm 2.5}$ | $50.4_{\pm 3.1}$ | $60.7_{\pm 1.7}$ | $78.6_{\pm 0.6}$ | $77.4_{\pm 0.9}$ | 69.3 |
| ASH[60] | $78.5_{\pm 3.2}$ | $47.3_{\pm 2.8}$ | $39.6_{\pm 1.7}$ | $78.0_{\pm 0.2}$ | **86.6**$_{\pm 0.7}$ | 66.0 |
| MDS[11] | $90.2_{\pm 0.1}$ | $57.8_{\pm 0.5}$ | $91.8_{\pm 0.1}$ | $62.9_{\pm 0.8}$ | $58.4_{\pm 0.1}$ | 72.2 |
| RMDS[46] | $59.4_{\pm 0.1}$ | $33.6_{\pm 1.4}$ | $47.4_{\pm 0.2}$ | $\underline{81.9}_{\pm 0.4}$ | $68.8_{\pm 0.1}$ | 58.2 |
| KNN[47] | $\underline{98.6}_{\pm 0.0}$ | $54.5_{\pm 0.5}$ | $91.1_{\pm 0.1}$ | $79.7_{\pm 0.6}$ | $77.4_{\pm 0.0}$ | $\underline{80.3}$ |
| SHE[48] | $88.3_{\pm 0.2}$ | $42.7_{\pm 0.6}$ | $73.2_{\pm 0.1}$ | $54.8_{\pm 0.7}$ | $83.0_{\pm 0.1}$ | 68.4 |
| NECO[51] | $53.5_{\pm 1.6}$ | $39.5_{\pm 3.2}$ | $35.1_{\pm 1.5}$ | $80.2_{\pm 0.1}$ | $67.2_{\pm 0.3}$ | 55.1 |
| NNGuide[49] | $70.6_{\pm 2.9}$ | $49.8_{\pm 4.2}$ | $43.6_{\pm 2.1}$ | $79.4_{\pm 0.0}$ | $85.1_{\pm 0.8}$ | 65.7 |
| Relation[50] | $80.7_{\pm 0.2}$ | $\underline{60.4}_{\pm 2.5}$ | $\underline{96.0}_{\pm 0.5}$ | $74.5_{\pm 0.3}$ | $81.8_{\pm 0.7}$ | 78.7 |
| SCALE[52] | $89.0_{\pm 2.9}$ | $44.9_{\pm 3.2}$ | $54.4_{\pm 2.1}$ | $78.4_{\pm 0.4}$ | $\underline{86.2}_{\pm 0.5}$ | 70.6 |
| fDBD[53] | $71.1_{\pm 0.5}$ | $51.3_{\pm 1.3}$ | $47.4_{\pm 0.2}$ | $79.9_{\pm 0.0}$ | $84.2_{\pm 0.3}$ | 66.8 |
| NCI[54] | $84.0_{\pm 0.1}$ | $46.4_{\pm 2.4}$ | $54.8_{\pm 0.8}$ | $78.5_{\pm 0.1}$ | $84.9_{\pm 0.2}$ | 69.7 |
| **SPROD** | **98.8**$_{\pm 0.0}$ | **61.6**$_{\pm 0.9}$ | **97.4**$_{\pm 0.0}$ | **82.4**$_{\pm 0.5}$ | $85.3_{\pm 0.0}$ | **85.1** |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP[7] | $87.9_{\pm 0.8}$ | $98.7_{\pm 0.5}$ | $97.3_{\pm 0.3}$ | $83.8_{\pm 0.7}$ | $74.1_{\pm 1.2}$ | 88.4 |
| Energy[13] | $89.2_{\pm 3.2}$ | $98.6_{\pm 0.7}$ | $95.5_{\pm 3.1}$ | $84.8_{\pm 0.8}$ | $76.3_{\pm 0.9}$ | 88.9 |
| MLS[44] | $88.1_{\pm 2.0}$ | $98.8_{\pm 0.6}$ | $96.7_{\pm 2.0}$ | $84.4_{\pm 0.8}$ | $74.6_{\pm 0.9}$ | 88.5 |
| KLM[44] | $89.1_{\pm 0.7}$ | $98.7_{\pm 0.5}$ | $97.1_{\pm 0.3}$ | $80.5_{\pm 0.8}$ | $76.1_{\pm 1.7}$ | 88.3 |
| GEN[45] | $87.9_{\pm 0.8}$ | $98.7_{\pm 0.5}$ | $97.3_{\pm 0.3}$ | $84.8_{\pm 0.1}$ | $76.3_{\pm 0.7}$ | 89.0 |
| GNorm[61] | $84.2_{\pm 0.7}$ | $98.8_{\pm 0.4}$ | $97.1_{\pm 0.1}$ | $84.2_{\pm 0.6}$ | $54.7_{\pm 0.6}$ | 83.8 |
| ReAct[58] | $86.9_{\pm 7.0}$ | $96.3_{\pm 2.4}$ | $95.5_{\pm 3.2}$ | $83.9_{\pm 0.8}$ | $57.5_{\pm 1.6}$ | 84.0 |
| VIM[59] | $61.4_{\pm 3.5}$ | $96.2_{\pm 0.4}$ | $69.0_{\pm 1.5}$ | $86.6_{\pm 0.7}$ | $79.5_{\pm 0.5}$ | 78.5 |
| ASH[60] | $85.2_{\pm 7.0}$ | $96.9_{\pm 1.4}$ | $96.1_{\pm 1.5}$ | $87.9_{\pm 0.4}$ | $52.9_{\pm 3.1}$ | 83.8 |
| MDS[11] | $49.2_{\pm 0.2}$ | $96.0_{\pm 0.5}$ | $39.0_{\pm 0.3}$ | $93.0_{\pm 0.3}$ | $90.5_{\pm 0.1}$ | 73.5 |
| RMDS[46] | $91.7_{\pm 0.2}$ | $99.6_{\pm 0.1}$ | $95.3_{\pm 0.1}$ | $83.4_{\pm 0.9}$ | $88.1_{\pm 0.1}$ | 91.6 |
| KNN[47] | $\underline{4.8}_{\pm 0.1}$ | $\underline{94.4}_{\pm 1.0}$ | $42.5_{\pm 0.2}$ | $79.9_{\pm 1.1}$ | $70.4_{\pm 0.2}$ | $\underline{58.4}$ |
| SHE[48] | $33.2_{\pm 0.5}$ | $96.4_{\pm 0.5}$ | $76.5_{\pm 0.2}$ | $93.9_{\pm 0.3}$ | $\underline{52.6}_{\pm 0.8}$ | 70.5 |
| NECO[51] | $90.5_{\pm 2.2}$ | $98.8_{\pm 0.6}$ | $96.7_{\pm 1.9}$ | $\underline{78.2}_{\pm 0.0}$ | $89.9_{\pm 0.8}$ | 90.8 |
| NNGuide[49] | $77.7_{\pm 6.0}$ | $97.6_{\pm 1.2}$ | $91.6_{\pm 3.6}$ | $86.3_{\pm 0.1}$ | **52.2**$_{\pm 1.4}$ | 81.1 |
| Relation[50] | $73.8_{\pm 0.5}$ | $95.4_{\pm 0.2}$ | $\underline{24.2}_{\pm 1.7}$ | $84.6_{\pm 0.2}$ | $78.0_{\pm 1.1}$ | 71.2 |
| SCALE[52] | $61.0_{\pm 22.5}$ | $98.7_{\pm 0.6}$ | $94.6_{\pm 3.1}$ | $87.7_{\pm 0.3}$ | $53.0_{\pm 1.5}$ | 79.0 |
| fDBD[53] | $85.5_{\pm 0.8}$ | $98.6_{\pm 0.5}$ | $96.1_{\pm 0.4}$ | $85.4_{\pm 0.2}$ | $70.8_{\pm 1.0}$ | 87.3 |
| NCI[54] | $41.1_{\pm 0.1}$ | $99.4_{\pm 0.3}$ | $92.2_{\pm 0.6}$ | $85.8_{\pm 0.3}$ | $63.8_{\pm 0.9}$ | 76.5 |
| **SPROD** | **4.7**$_{\pm 0.1}$ | **93.7**$_{\pm 0.9}$ | **19.0**$_{\pm 0.4}$ | **69.5**$_{\pm 1.2}$ | $58.0_{\pm 0.1}$ | **49.0** |




(a) ResNet-50 backbone.

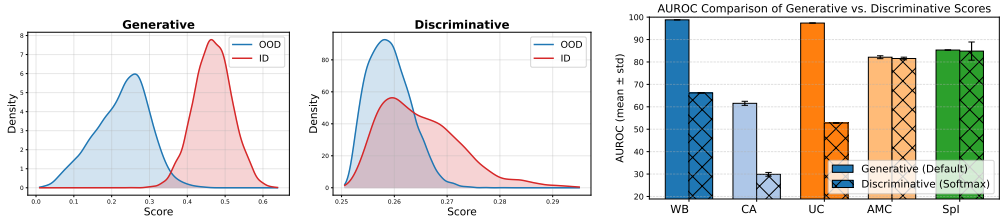



(b) ResNet-18 backbone.

Figure 4: Effect of backbone fine-tuning and spurious correlation on SP-OOD detection using the Waterbirds dataset. Left: ResNet-50; right: ResNet-18. Each pair shows results under 50% (left) and 90% (right) spurious correlation in ID data. Fine-tuned models are marked with a hatch texture.

### 5.3 Results

The primary results, summarized in Table 2, demonstrate the efficacy of SPROD across various datasets using a ResNet-50 backbone, evaluated with both AUROC and FPR@95 metrics. Additional results on other backbone architectures are provided in Appendix G. SPROD consistently achieves superior performance compared to the 19 post-hoc baselines in all conducted experiments. In contrast, competing methods exhibit more variable performance, excelling in only a subset of the experimental settings. Generally, feature-based OOD detection methods tend to outperform output-based methods; however, this advantage appears to diminish on larger-scale datasets with multiple classes. Overall, the average performance across all datasets reveals a notable margin by which SPROD surpasses the other evaluated baselines: Specifically, SPROD reaches an average of 85.1% AUROC, 4.8% higher (absolute) than the second best, KNN. For FPR@95, SPROD reaches 49.1% error rate on average, 9.3% better than the second best, KNN. All other compared approaches are more than 20% worse.




(a) (b)

Figure 5: Comparison of generative and discriminative scoring for OOD detection using SPROD. (a) Histograms of ID and OOD sample scores using the distance-based generative approach and the softmax-based discriminative approach, both computed with SPROD on the Waterbirds dataset. (b) Performance comparison between the generative (distance-based) and discriminative (softmax-based) scoring variants of SPROD across the five SP-OOD benchmark datasets.

Table 3: Comparison with the best-performing methods reported in the OpenOOD [37] on conventional (standard) OOD datasets (AUROC↑). The highest scores are highlighted in **bold**.

| Method | CIFAR-10 Near | CIFAR-10 Far | CIFAR-100 Near | CIFAR-100 Far | ImageNet Near | ImageNet Far | Avg (%) |
|---|---|---|---|---|---|---|---|
| RMDS [46] | 89.80 | 92.20 | 80.15 | **82.92** | 76.99 | 86.38 | 84.74 |
| MLS [44] | 87.52 | 91.10 | 81.05 | 79.67 | 76.46 | 89.57 | 84.23 |
| VIM [59] | 88.68 | **93.48** | 74.98 | 81.70 | 72.08 | 92.68 | 83.27 |
| KNN [47] | **90.64** | 92.96 | 80.18 | 82.40 | 71.10 | 90.18 | 84.58 |
| ASH [60] | 75.12 | 78.49 | 78.20 | 80.58 | **78.17** | **95.74** | 81.05 |
| **SPROD** | 89.04 | 91.78 | **81.80** | 79.93 | 75.06 | 95.29 | **85.15** |

Next, we investigate the impact of backbone fine-tuning and correlation rate on OOD detection performance, using the Waterbirds dataset with ResNet-50 (Figure 4a) and ResNet-18 (Figure 4b) backbones. Fine-tuning the backbone on this dataset generally degrades OOD detection performance, a finding that contrasts with some common assumptions. This negative effect appears to be more pronounced for feature-based methods. As expected, increasing the spurious correlation rate of ID data from 50% to 90% leads to a general decline in performance across methods; however, this degradation is more noticeable for output-based techniques. Furthermore, the results suggest that employing a lighter, less expressive backbone (ResNet-18 compared to ResNet-50) does not lead to a substantial performance drop in OOD detection. Across these variations, SPROD demonstrates consistent robustness and maintains its performance as the spurious correlation rate increases.

To investigate the impact of the scoring mechanism, as discussed in Section 3.2, we conduct an ablation study using SPROD as the base method. This controlled experiment compares the effectiveness of deriving OOD scores in a discriminative manner $p(y|z)$ versus a generative manner $p(z|y)$. Both baselines utilize the same samples, feature embeddings, and refined prototypes from SPROD. For the discriminative score, we apply a softmax function to the negative distances between a sample's embedding $z$ and the class prototypes. The default SPROD approach, which uses the negative of the minimum distance to class prototypes, serves as the generative baseline (aligning with a log-likelihood under an exponential family distribution assumption).

Figure 5a presents histograms of the scores generated by both approaches, showing that the generative scoring method yields more distinctly separated distributions for ID and OOD samples. Furthermore, Figure 5b compares the OOD detection performance of these two scoring variants across our five SP-OOD benchmark settings. The results indicate that applying the softmax function for discriminative scoring substantially degrades performance, particularly in WaterBirds, CelebA and UrbanCars datasets. Conversely, on Animals MetaCoCo and Spurious ImageNet, the performance degradation from using softmax is less pronounced. This observation aligns with the trends in Table 2, where traditional output-based methods tend to perform relatively better in these datasets.

To further assess the generality of SPROD beyond spurious correlation settings, we evaluate it on conventional OOD benchmarks using the standardized OpenOOD protocol [37]. We compare SPROD with the top-performing post-hoc methods reported in the OpenOOD study. Table 3 shows that SPROD attains performance comparable to or exceeding these methods and achieves the highest average score across all settings, confirming that SPROD performs reliably on standard OOD benchmarks and is not limited to spurious correlation scenarios.

While the simplicity of its prototypical framework makes SPROD a computationally efficient post-hoc method, its sample efficiency in SP-OOD settings also deserves investigation. To this end, we evaluate SPROD alongside the second top-performing method from Table 2 under low-shot conditions. In this setup, we reduce the number of ID training samples while carefully preserving the original level of spurious correlation present in the experimental setting. The results, presented in Figure 6, demonstrate that both SPROD and KNN maintain strong performance even in this data-scarce regime, highlighting the sample efficiency of these post-hoc approaches. Interestingly, for the CelebA dataset, both methods exhibit improved OOD detection performance in the low-shot setting compared to when trained on the full dataset.

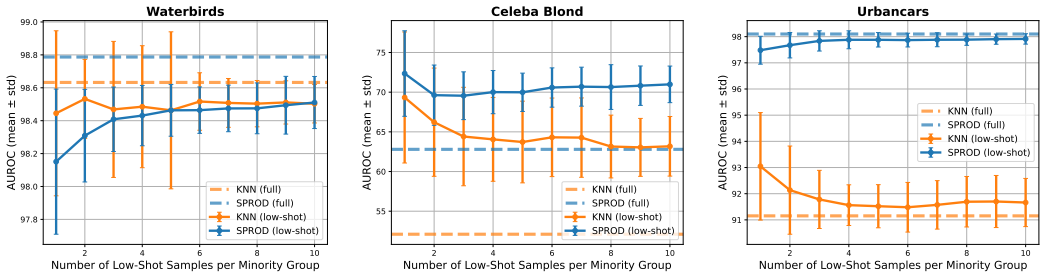



Figure 6: Performance of SPROD and KNN in low-shot SP-OOD settings across three datasets. Dashed lines indicate performance with the full training set, while solid lines show performance using varying numbers of samples per minority group.

Zero-shot OOD detection methods have reported results on Waterbirds using CLIP-B/16, but since they operate in a zero-shot setting, spurious correlations in the training set are not meaningful for them. Still, we compare: MCM [29] and CMA [32] achieve 98.36 and 98.62 AUROC using text inputs, while our text-free, vision-only method outperforms both with 99.01. COVER [28] also evaluates this setting but over a different operating range, achieving lower AUROC scores (90.52 vs. 90.31 for MCM), highlighting our method's superior performance. Additional comparisons with existing SP-OOD methods (specifically CLIP-based approaches) are provided in Appendix C.

## 6 Discussion

This paper introduces SPROD, a prototype-based method enhancing out-of-distribution (OOD) detection robustness against unknown spurious correlations. SPROD refines class prototypes by identifying and then adjusting for potential subgroups influenced by spurious features, thereby aligning representations with core, invariant class characteristics. A key strength is SPROD's efficiency and adaptability: as a post-hoc method, it integrates with various pre-trained feature extraction models without requiring retraining, additional OOD data, or hyperparameter tuning.

Experimental results consistently demonstrate the superiority of SPROD across ten convolutional and transformer-based backbones and five diverse SP-OOD benchmarks, including the newly introduced Animals MetaCoCo dataset. Our evaluations also reveal that fine-tuning the feature backbone on ID data can degrade SP-OOD detection performance. Furthermore, investigations into scoring mechanisms highlighted the advantages of distance-based approaches over softmax-based scoring for SP-OOD detection, particularly affirming the design choices in SPROD.

SPROD offers a scalable solution for improving robustness in OOD detection. While demonstrating significant advancements, SPROD's current formulation relies on class labels from the ID training data to construct class-conditional prototypes. This reliance aligns with the typical assumptions of the spurious correlation setting, which presumes access to data samples with spuriously correlated labels. SPROD is not always the single best-performing method on every metric or in every setting (particularly in the conventional setting), but it remains competitive with the best-performing approaches. Such trade-offs are common in robustness research, where minor reductions in overall accuracy are often accepted to attain higher worst-group performance. Future work could explore theoretical justifications for the robustness of the generative-like scoring mechanism or investigate more expressive approaches for modeling class-conditional distributions.

## Acknowledgments

The research at TU Darmstadt was partially funded by an **Alexander von Humboldt Professorship in Multimodal Reliable AI**, sponsored by Germany's Federal Ministry for Education and Research.

## References

[1] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images, 2015. URL https://arxiv.org/abs/1412.1897.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR*, abs/1502.01852, 2015. URL http://arxiv.org/abs/1502.01852.

[3] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, may 2017. ISSN 0001-0782. doi: 10.1145/3065386. URL https://doi.org/10.1145/3065386.

[4] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *CoRR*, abs/1804.02767, 2018. URL http://arxiv.org/abs/1804.02767.

[5] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. URL http://arxiv.org/abs/1506.01497.

[6] Anh Mai Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CoRR*, abs/1412.1897, 2014. URL http://arxiv.org/abs/1412.1897.

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *CoRR*, abs/1610.02136, 2016. URL http://arxiv.org/abs/1610.02136.

[8] David Zimmerer, Peter Full, Fabian Isensee, Paul Jager, Tim Adler, Jens Petersen, Gregor Kohler, Tobias Roß, Annika Reinke, Antanas Kascenas, Bjorn Jensen, Alison O'Neil, Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, Bernhard Kainz, Nina Shvetsova, Irina Fedulova, and Klaus Maier-Hein. Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging*, PP:1–1, 04 2022. doi: 10.1109/TMI.2022.3170077.

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. doi: 10.1109/CVPR.2012.6248074.

[10] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks, 2020. URL https://arxiv.org/abs/1706.02690.

[11] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks, 2018. URL https://arxiv.org/abs/1807.03888.

[12] Dan Hendrycks, Mantas Mazeika, and Thomas G. Dietterich. Deep anomaly detection with outlier exposure. *CoRR*, abs/1812.04606, 2018. URL http://arxiv.org/abs/1812.04606.

[13] 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.

[14] Yifei Ming, Hang Yin, and Yixuan Li. On the impact of spurious correlation for out-of-distribution detection. *CoRR*, abs/2109.05642, 2021. URL https://arxiv.org/abs/2109.05642.

[15] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *Proceedings of the European conference on computer vision (ECCV)*, pages 456–473, 2018.

[16] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

[17] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731, 2019. URL http://arxiv.org/abs/1911.08731.

[18] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *CoRR*, abs/2004.07780, 2020. URL https://arxiv.org/abs/2004.07780.

[19] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 2022. URL https://arxiv.org/abs/2204.02937.

[20] Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information, 2021. URL https://arxiv.org/abs/2107.09044.

[21] Rundong He, Zhongyi Han, Xiankai Lu, and Yilong Yin. Ronf: Reliable outlier synthesis under noisy feature space for out-of-distribution detection. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4242–4251, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3547815. URL https://doi.org/10.1145/3503161.3547815.

[22] Jaeyoung Kim, Seo Taek Kong, Dongbin Na, and Kyu-Hwan Jung. Key feature replacement of in-distribution samples for out-of-distribution detection, 2022. URL https://arxiv.org/abs/2301.13012.

[23] Kai Liu, Zhihang Fu, Sheng Jin, Chao Chen, Ze Chen, Rongxin Jiang, Fan Zhou, Yaowu Chen, and Jieping Ye. Rethinking out-of-distribution detection on imbalanced data distribution, 2024. URL https://arxiv.org/abs/2407.16430.

[24] Yu Wang, Junxian Mu, Hongzhi Huang, Qilong Wang, Pengfei Zhu, and Qinghua Hu. Backmix: Regularizing open set recognition by removing underlying fore-background priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12, 2025. doi: 10.1109/TPAMI.2025.3550703.

[25] Lily H. Zhang and Rajesh Ranganath. Robustness to spurious correlations improves semantic out-of-distribution detection, 2023.

[26] Zhen Fang, Jie Lu, and Guangquan Zhang. Out-of-distribution detection with non-semantic exploration. *Information Sciences*, 705:121989, 2025. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2025.121989. URL https://www.sciencedirect.com/science/article/pii/S0020025525001215.

[27] Sungik Choi, Hankook Lee, Honglak Lee, and Moontae Lee. Projection regret: Reducing background bias for novelty detection via diffusion models, 2023. URL https://arxiv.org/abs/2312.02615.

[28] Boxuan Zhang, Jianing Zhu, Zengmao Wang, Tongliang Liu, Bo Du, and Bo Han. What if the input is expanded in ood detection?, 2024. URL https://arxiv.org/abs/2410.18472.

[29] Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations, 2022.

[30] Yi Dai, Hao Lang, Kaisheng Zeng, Fei Huang, and Yongbin Li. Exploring large language models for multi-modal out-of-distribution detection. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5292–5305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/

v1/2023.findings-emnlp.351. URL https://aclanthology.org/2023.findings-emnlp.351.

[31] Xue Jiang, Feng Liu, Zhen Fang, Hong Chen, Tongliang Liu, Feng Zheng, and Bo Han. Negative label guided ood detection with pretrained vision-language models, 2024. URL https://arxiv.org/abs/2403.20078.

[32] Yuxiao Lee, Xiaofeng Cao, Jingcai Guo, Wei Ye, Qing Guo, and Yi Chang. Concept matching with agent for out-of-distribution detection, 2025. URL https://arxiv.org/abs/2405.16766.

[33] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization, 2020. URL https://arxiv.org/abs/1911.08731.

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[35] Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Canton Ferrer, Chenliang Xu, and Mark Ibrahim. A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20071–20082, 2023.

[36] Yannic Neuhaus, Maximilian Augustin, Valentyn Boreiko, and Matthias Hein. Spurious features everywhere – large-scale detection of harmful spurious features in imagenet, 2023. URL https://arxiv.org/abs/2212.04871.

[37] Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

[38] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9518–9526, 2019.

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

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

[41] Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019.

[42] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021.

[43] Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pages 23631–23644. PMLR, 2022.

[44] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. A benchmark for anomaly segmentation. *CoRR*, abs/1911.11132, 2019. URL http://arxiv.org/abs/1911.11132.

[45] Xixi Liu, Yaroslava Lochman, and Christopher Zach. Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23946–23955, 2023.

[46] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection, 06 2021.

[47] 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.

[48] Jinsong Zhang, Qiang Fu, Xu Chen, Lun Du, Zelin Li, Gang Wang, Shi Han, Dongmei Zhang, et al. Out-of-distribution detection based on in-distribution data patterns memorization with modern hopfield energy. In *The Eleventh International Conference on Learning Representations*, 2022.

[49] Jaewoo Park, Yoon Gyo Jung, and Andrew Beng Jin Teoh. Nearest neighbor guidance for out-of-distribution detection, 2023. URL https://arxiv.org/abs/2309.14888.

[50] Jang-Hyun Kim, Sangdoo Yun, and Hyun Oh Song. Neural relation graph: A unified framework for identifying label noise and outlier data, 2023. URL https://arxiv.org/abs/2301.12321.

[51] Mouïn Ben Ammar, Nacim Belkhir, Sebastian Popescu, Antoine Manzanera, and Gianni Franchi. Neco: Neural collapse based out-of-distribution detection. *arXiv preprint arXiv:2310.06823*, 2023.

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

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

[54] 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.

[55] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection?, 2023. URL https://arxiv.org/abs/2203.04450.

[56] Haodong Lu, Dong Gong, Shuo Wang, Jason Xue, Lina Yao, and Kristen Moore. Learning with mixture of prototypes for out-of-distribution detection, 2024. URL https://arxiv.org/abs/2402.02653.

[57] Poulami Sinhamahapatra, Franziska Schwaiger, Shirsha Bose, Huiyu Wang, Karsten Roscher, and Stephan Günnemann. Finding dino: A plug-and-play framework for zero-shot detection of out-of-distribution objects using prototypes. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8474–8483. IEEE, 2025.

[58] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *CoRR*, abs/2111.12797, 2021. URL https://arxiv.org/abs/2111.12797.

[59] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching, 2022. URL https://arxiv.org/abs/2203.10807.

[60] Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection, 2023. URL https://arxiv.org/abs/2209.09858.

[61] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *CoRR*, abs/2110.00218, 2021. URL https://arxiv.org/abs/2110.00218.

[62] Chirag Agarwal and Sara Hooker. Estimating example difficulty using variance of gradients. *CoRR*, abs/2008.11600, 2020. URL https://arxiv.org/abs/2008.11600.

[63] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with in-distribution examples and gram matrices. *CoRR*, abs/1912.12510, 2019. URL http://arxiv.org/abs/1912.12510.

[64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[65] Mohammadamin Banayeeanzade, Rasoul Mirzaiezadeh, Hosein Hasani, and Mahdieh Soleymani. Generative vs. discriminative: Rethinking the meta-continual learning. *Advances in Neural Information Processing Systems*, 34:21592–21604, 2021.

[66] Alexander Cong Li, Ananya Kumar, and Deepak Pathak. Generative classifiers avoid shortcut solutions. In *The Thirteenth International Conference on Learning Representations*, 2024.

[67] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[68] 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 (CVPR)*, June 2016.

[69] 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, 2018. doi: 10.1109/TPAMI.2017.2723009.

[70] Bryson Lingenfelter, Sara R. Davis, and Emily M. Hand. A quantitative analysis of labeling issues in the celeba dataset. In George Bebis, Bo Li, Angela Yao, Yang Liu, Ye Duan, Manfred Lau, Rajiv Khadka, Ana Crisan, and Remco Chang, editors, *Advances in Visual Computing*, pages 129–141, Cham, 2022. Springer International Publishing. ISBN 978-3-031-20713-6.

[71] Min Zhang, Haoxuan Li, Fei Wu, and Kun Kuang. Metacoco: A new few-shot classification benchmark with spurious correlation, 2024. URL https://arxiv.org/abs/2404.19644.

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

[73] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2014. URL https://arxiv.org/abs/1409.0575.

[74] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Cub dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[75] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 554–561, 2013. doi: 10.1109/ICCVW.2013.77.

[76] Agrim Gupta, Piotr Dollár, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation, 2019. URL https://arxiv.org/abs/1908.03195.

[77] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128 (7):1956–1981, March 2020. ISSN 1573-1405. doi: 10.1007/s11263-020-01316-z. URL http://dx.doi.org/10.1007/s11263-020-01316-z.

[78] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Ng. Reading digits in natural images with unsupervised feature learning. *NIPS*, 01 2011.

[79] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015. URL http://arxiv.org/abs/1506.03365.

[80] Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R. Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking, 2015. URL https://arxiv.org/abs/1504.06755.

[81] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[82] 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, 2021. URL https://arxiv.org/abs/2010.11929.

[83] 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, 2021. URL https://arxiv.org/abs/2103.14030.

[84] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s, 2022. URL https://arxiv.org/abs/2201.03545.

[85] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Large scale learning of general visual representations for transfer. *CoRR*, abs/1912.11370, 2019. URL http://arxiv.org/abs/1912.11370.

[86] Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection, 2023.

[87] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *CoRR*, abs/2110.11334, 2021. URL https://arxiv.org/abs/2110.11334.

# A Theoretical Analysis of SPROD

SPROD employs a two-step prototype refinement strategy to approximate group-specific representations within each class. The first step creates new prototypes for misclassified training samples, motivated by the observation that minority group instances (those that deviate from dominant spurious patterns) are more prone to misclassification. However, some correctly classified minority samples may still be incorrectly assigned to biased prototypes.

To address this, the second step reassigns each training sample to its nearest prototype and recalculates prototypes based on these updated assignments. This reassignment helps produce cleaner subgroup representations by reducing the influence of spurious-majority samples on minority feature structure.

By applying these two steps, SPROD aims to obtain nearly pure group-specific prototypes. In the remainder of this section, we provide a theoretical formulation to explain why such subgroup-specific prototypes improve robustness to spurious OOD (SP-OOD) samples.

### A.0.1 Feature Decomposition

Without loss of generality, we assume a binary classification problem with classes $c \in \{0, 1\}$. Each input sample $x_i$ is mapped to a normalized feature embedding $z_i \in \mathbb{R}^d$. We consider a general decomposition of $z_i$ into three semantically distinct components: spurious, core, and irrelevant features. We model $z_i$ as a linear combination of three functionally distinct components that span the embedding space:

$$z_i = \sum_{j=1}^{n_{u_c^{\text{sp}}}} \alpha_{i,j}^{\text{sp}} \, \vec{\mathbf{u}}_{c,j}^{\text{sp}} + \sum_{j=1}^{n_{u_c^{\text{core}}}} \beta_{i,j}^{\text{core}} \, \vec{\mathbf{u}}_{c,j}^{\text{core}} + \sum_{j=1}^{n_{u^{\text{irr}}}} \gamma_{i,j}^{\text{irr}} \, \vec{\mathbf{u}}_{j}^{\text{irr}},$$

where the sets $\{\vec{\mathbf{u}}_{c,j}^{\text{sp}}\}$, $\{\vec{\mathbf{u}}_{c,j}^{\text{core}}\}$, and $\{\vec{\mathbf{u}}_{j}^{\text{irr}}\}$ form orthonormal bases for the spurious, core, and irrelevant subspaces, respectively. Each set of coefficients $\{\alpha_{i,j}^{\text{sp}}\}$, $\{\beta_{i,j}^{\text{core}}\}$, and $\{\gamma_{i,j}^{\text{irr}}\}$ is specific to instance $i$.

These subspaces span mutually exclusive semantic roles:

- Spurious subspace: captures features correlated with the label during training but not causally predictive.
- Core subspace: captures features that are causally predictive of the label.
- Irrelevant subspace: captures features unrelated to the task.

By definition, the basis vectors across subspaces are orthogonal: if an irrelevant basis vector were correlated with a core basis, it would be predictive and thus belong in the core subspace.

This decomposition can be expressed compactly in matrix form:

$$z_i = U_c^{\text{sp}} \, \vec{\boldsymbol{\alpha}}_i^{\text{sp}} + U_c^{\text{core}} \, \vec{\boldsymbol{\beta}}_i^{\text{core}} + U^{\text{irr}} \, \vec{\boldsymbol{\gamma}}_i^{\text{irr}},$$

where each $U^{(\cdot)} \in \mathbb{R}^{d \times n_u^{(\cdot)}}$ is a matrix of orthonormal basis vectors for the corresponding subspace, and each coefficient vector (e.g., $\vec{\boldsymbol{\alpha}}_i^{\text{sp}} \in \mathbb{R}^{n_{u_c^{\text{sp}}}}$) represents the coordinates of $z_i$ in that subspace.

### A.0.2 Distributional Assumptions

We assume the coefficient vectors for each instance $z_i$ are drawn from class-conditional distributions:

$$\vec{\boldsymbol{\alpha}}_i^{\text{sp}} \sim P_c^{\text{sp}}, \quad \vec{\boldsymbol{\beta}}_i^{\text{core}} \sim P_c^{\text{core}},$$

where $P_c^{\text{sp}}$ and $P_c^{\text{core}}$ are distributions over the spurious and core subspace coefficients for class $c$.

The irrelevant component is shared across classes:

$$\vec{\boldsymbol{\gamma}}_i^{\text{irr}} \sim P^{\text{irr}}.$$

### A.0.3 Group Definitions

Within each class $c$, we define majority and minority groups based on the alignment of spurious features.

**Majority group** samples $z_i \in \mathcal{S}_c^{\text{maj}}$ satisfy:

$$\vec{\boldsymbol{\alpha}}_i^{\text{sp}} \sim P_c^{\text{sp}}, \quad \vec{\boldsymbol{\beta}}_i^{\text{core}} \sim P_c^{\text{core}}.$$

**Minority group** samples $z_i \in \mathcal{S}_c^{\text{min}}$ satisfy:

$$\vec{\boldsymbol{\alpha}}_i^{\text{sp}} \sim P_{1-c}^{\text{sp}}, \quad \vec{\boldsymbol{\beta}}_i^{\text{core}} \sim P_c^{\text{core}}.$$

Let $C_c$ denote the number of samples in class $c$, and $|\mathcal{S}_c^{\text{maj}}|$ the number of majority samples. Then:

$$r_c = \frac{|\mathcal{S}_c^{\text{maj}}|}{C_c} \in [0, 1]$$

quantifies the class-conditional spurious correlation strength.

### A.0.4 OOD Sample Definition

OOD samples may contain components beyond the core, spurious, and irrelevant subspaces. To model this, we extend the feature decomposition to include an additional subspace $U^{\text{ext}}$, capturing external factors not observed during training. We define an OOD sample as:

$$z_{\text{OOD}} = U_c^{\text{sp}} \vec{\boldsymbol{\alpha}}_{\text{OOD}}^{\text{sp}} + U_c^{\text{irr}} \vec{\boldsymbol{\gamma}}_{\text{OOD}}^{\text{irr}} + U^{\text{ext}} \vec{\boldsymbol{\delta}}_{\text{OOD}}^{\text{ext}},$$

where $\vec{\boldsymbol{\alpha}}_{\text{OOD}}^{\text{sp}} \sim P_c^{\text{sp}}$, $\vec{\boldsymbol{\gamma}}_{\text{OOD}}^{\text{irr}} \sim P^{\text{irr}}$, and $\vec{\boldsymbol{\delta}}_{\text{OOD}}^{\text{ext}}$ is unconstrained. Hard OOD samples may also include core components drawn from a shifted distribution $\vec{\boldsymbol{\beta}}_{\text{OOD}}^{\text{core}} \sim Q^{\text{core}} \neq P_c^{\text{core}}$. Although it is possible to analyze the general form with core and external components, for simplicity we focus on near-OOD samples with no core or external components, i.e.,

$$z_{\text{OOD}} = U_c^{\text{sp}} \vec{\boldsymbol{\alpha}}_{\text{OOD}}^{\text{sp}} + U_c^{\text{irr}} \vec{\boldsymbol{\gamma}}_{\text{OOD}}^{\text{irr}}, \quad \text{with } \vec{\boldsymbol{\beta}}_{\text{OOD}}^{\text{core}} = \vec{0},\ \vec{\boldsymbol{\delta}}_{\text{OOD}}^{\text{ext}} = \vec{0}.$$

We define spurious-OOD groups based on the source of the spurious component:

$$\mathcal{S}_c^{\text{OOD}} = \left\{ z_{\text{OOD}} \,\middle|\, \vec{\boldsymbol{\alpha}}_{\text{OOD}}^{\text{sp}} \sim P_c^{\text{sp}},\ \vec{\boldsymbol{\gamma}}_{\text{OOD}}^{\text{irr}} \sim P^{\text{irr}} \right\}, \quad c \in \{0, 1\}.$$

These near-OOD samples resemble class $c$ only through spurious features and match the in-distribution (ID) irrelevant component distribution.

### A.0.5 Prototype Calculation

We define the expected coefficient vectors for each subspace:

$$\vec{\boldsymbol{\mu}}_c^{\text{core}} = \mathbb{E}_{\vec{\boldsymbol{\beta}} \sim P_c^{\text{core}}}[\vec{\boldsymbol{\beta}}], \quad \vec{\boldsymbol{\mu}}_c^{\text{sp}} = \mathbb{E}_{\vec{\boldsymbol{\alpha}} \sim P_c^{\text{sp}}}[\vec{\boldsymbol{\alpha}}], \quad \vec{\boldsymbol{\mu}}^{\text{irr}} = \mathbb{E}_{\vec{\boldsymbol{\gamma}} \sim P^{\text{irr}}}[\vec{\boldsymbol{\gamma}}].$$

Using these, we define the majority and minority subgroup prototypes for class $c \in \{0, 1\}$ as:

$$\vec{\mathbf{p}}_c^{\text{maj}} = U_c^{\text{sp}} \vec{\boldsymbol{\mu}}_c^{\text{sp}} + U_c^{\text{core}} \vec{\boldsymbol{\mu}}_c^{\text{core}} + U^{\text{irr}} \vec{\boldsymbol{\mu}}^{\text{irr}},$$

$$\vec{\mathbf{p}}_c^{\text{min}} = U_c^{\text{sp}} \vec{\boldsymbol{\mu}}_{1-c}^{\text{sp}} + U_c^{\text{core}} \vec{\boldsymbol{\mu}}_c^{\text{core}} + U^{\text{irr}} \vec{\boldsymbol{\mu}}^{\text{irr}}.$$

The overall class prototype is a convex combination of subgroup prototypes:

$$\vec{\mathbf{p}}_c = r_c \vec{\mathbf{p}}_c^{\text{maj}} + (1 - r_c)\vec{\mathbf{p}}_c^{\text{min}},$$

where $r_c \in [0, 1]$ denotes the proportion of majority group samples in class $c$.

### A.0.6 Bias in Prototype Distances Under Strong Spurious Correlation

In standard prototype-based OOD detection, the class prototype $\vec{\mathbf{p}}_c$ is a convex combination of majority and minority subgroup prototypes. In spurious-dominated regimes where $r_c \approx 1$, we have:

$$\vec{\mathbf{p}}_c \approx \vec{\mathbf{p}}_c^{\text{maj}} = U_c^{\text{sp}} \vec{\boldsymbol{\mu}}_c^{\text{sp}} + U_c^{\text{core}} \vec{\boldsymbol{\mu}}_c^{\text{core}} + U^{\text{irr}} \vec{\boldsymbol{\mu}}^{\text{irr}}.$$

**Case 1: Spurious-OOD Sample.** Let $z_{\text{OOD}} \in \mathcal{S}_c^{\text{OOD}}$ be an OOD sample with spurious alignment to class $c$, but no core features:

$$z_{\text{OOD}} = U_c^{\text{sp}} \vec{\boldsymbol{\alpha}}_{\text{OOD}} + U^{\text{irr}} \vec{\boldsymbol{\gamma}}_{\text{OOD}}, \quad \text{with } \vec{\boldsymbol{\alpha}}_{\text{OOD}} \sim P_c^{\text{sp}},\ \vec{\boldsymbol{\gamma}}_{\text{OOD}} \sim P^{\text{irr}}.$$

The expected squared distance to the biased class prototype becomes:

$$\mathbb{E}\left[\|z_{\text{OOD}} - \vec{\mathbf{p}}_c\|^2\right] = \underbrace{\|\vec{\boldsymbol{\mu}}_c^{\text{core}}\|^2}_{\text{core bias}} + \underbrace{\text{Tr}(\Sigma_c^{\text{sp}})}_{\text{spurious variance}} + \underbrace{\text{Tr}(\Sigma^{\text{irr}})}_{\text{irrelevant variance}}.$$

**Case 2: Minority In-Distribution Sample.** Now consider an ID sample from the minority group:

$$z_{\text{min}} = U_c^{\text{sp}} \vec{\boldsymbol{\alpha}}_{\text{min}} + U_c^{\text{core}} \vec{\boldsymbol{\beta}}_{\text{min}} + U^{\text{irr}} \vec{\boldsymbol{\gamma}}_{\text{min}},$$

with:

$$\vec{\boldsymbol{\alpha}}_{\text{min}} \sim P_{1-c}^{\text{sp}}, \quad \vec{\boldsymbol{\beta}}_{\text{min}} \sim P_c^{\text{core}}, \quad \vec{\boldsymbol{\gamma}}_{\text{min}} \sim P^{\text{irr}}.$$

Its expected squared distance to the majority prototype is:

$$\mathbb{E}\left[\|z_{\text{min}} - \vec{\mathbf{p}}_c^{\text{maj}}\|^2\right] = \underbrace{\|\vec{\boldsymbol{\mu}}_{1-c}^{\text{sp}} - \vec{\boldsymbol{\mu}}_c^{\text{sp}}\|^2}_{\text{spurious bias}} + \underbrace{\text{Tr}(\Sigma_c^{\text{core}})}_{\text{core variance}} + \underbrace{\text{Tr}(\Sigma^{\text{irr}})}_{\text{irrelevant variance}}.$$

**Key Insight:** Although $z_{\text{min}}$ is an ID sample, its distance to the biased prototype includes a *spurious bias* term, while the OOD sample differs only in the *core* direction. Depending on the relative magnitudes of the core and spurious bias terms, this highlights the potential for erroneous OOD detection when prototype estimates are biased toward spurious features.

For example, in the Waterbird dataset, the background (water or land) represents the spurious feature dimension characterized by distinct mean vectors $\vec{\boldsymbol{\mu}}_c^{\text{sp}}$ and $\vec{\boldsymbol{\mu}}_{1-c}^{\text{sp}}$. The difference $\|\vec{\boldsymbol{\mu}}_c^{\text{sp}} - \vec{\boldsymbol{\mu}}_{1-c}^{\text{sp}}\|$ may exceed the core bias in OOD distances, causing minority samples with conflicting backgrounds to lie farther from prototypes than some OOD samples.

### A.0.7 Why SPROD Mitigates Distance Bias

If prototypes can be accurately estimated to match each group's distribution, then each ID sample $z_i \in \mathcal{S}_c^g$ can be compared to its corresponding prototype $\hat{\mathbf{p}}(z_i) = \vec{\mathbf{p}}_c^g$, where $g \in \{\text{maj}, \text{min}\}$.

Because the prototype shares the same spurious basis alignment, the spurious bias is eliminated. The expected squared distance becomes:

$$\mathbb{E}\left[\|z_i - \hat{\mathbf{p}}(z_i)\|^2\right] = \text{Tr}(\Sigma_c^{\text{core}}) + \text{Tr}(\Sigma^{\text{irr}}).$$

In contrast, the standard prototype-based approach introduces a systematic bias for minority samples:

$$\mathbb{E}\left[\|z_{\text{min}} - \vec{\mathbf{p}}_c^{\text{maj}}\|^2\right] = \|\vec{\boldsymbol{\mu}}_{1-c}^{\text{sp}} - \vec{\boldsymbol{\mu}}_c^{\text{sp}}\|^2 + \text{Tr}(\Sigma_c^{\text{core}}) + \text{Tr}(\Sigma^{\text{irr}}).$$

For an SP-OOD sample, the prototype it is compared against (either group) contains a core component it lacks. Hence the expected distance becomes:

$$\mathbb{E}\left[\|z_{\text{OOD}} - \hat{\mathbf{p}}\|^2\right] = \|\vec{\boldsymbol{\mu}}_c^{\text{core}}\|^2 + \text{Tr}(\Sigma_c^{\text{core}}) + \text{Tr}(\Sigma^{\text{irr}}).$$

This is strictly greater than the refined ID distance, which includes neither the spurious bias nor the core bias term:

$$\mathbb{E}\left[\|z_{\text{OOD}} - \hat{\mathbf{p}}\|^2\right] > \mathbb{E}\left[\|z_i - \hat{\mathbf{p}}(z_i)\|^2\right].$$

**Conclusion:** SPROD removes the spurious bias term for ID samples and reduces overlap with OOD samples by accounting for their missing core features. This systematic (in expectation) bias elimination leads to tighter ID clusters and more reliable OOD detection.

## B Spurious Correlation in OOD Detection: Literature Review

This work [14] was the first to introduce the problem of spurious correlations in OOD detection, showing that detection performance degrades as spurious correlation increases, especially for SP-OOD samples. It also encouraged the community to report results of OOD methods under this challenging setting. Recently, several works have addressed the challenge of **spurious correlations** in *out-of-distribution (OOD)* detection, employing different strategies that can be broadly categorized into *outlier exposure based*, *training-time regularization*, and *post-hoc methods*.

**Outlier exposure based methods** include **RONF** [21], which improves synthetic outlier generation and model fine-tuning using only ID data. It introduces *Boundary Feature Mixup* to create more realistic virtual outliers by interpolating near decision boundaries, and *Optimal Parameter Learning* to suppress spurious feature learning during training. At inference, it uses a custom *Energy with Energy Discrepancy* score to better separate ID from OOD samples without relying on external OOD datasets. Similarly, **KIRBY** [22] generates hard negative samples by removing class-discriminative regions identified via Class Activation Maps (CAM) and inpainting these regions with background-like content, creating semantically degraded but visually plausible images as surrogate OOD examples. A lightweight rejection network trained on features from both clean and modified images enables strong OOD detection without requiring real OOD data or backbone retraining. Although not directly targeting spurious correlations, **ImOOD** [23] evaluates robustness under spurious settings, focusing on long-tailed datasets where class imbalance biases OOD detection toward frequent (head) classes. It learns a bias correction term to shift OOD scores per input, improving separation especially for rare (tail) classes.

**Training-time regularization methods** seek to reduce reliance on spurious cues during model training. **BackMix** [24] regularizes models by mixing foreground objects with different backgrounds, breaking spurious correlations between objects and backgrounds. Using CAM to estimate foreground regions, it replaces background patches with those from other images while preserving labels, thus improving robustness primarily against background spuriousity. **RW** [25] introduces a nuisance-aware training framework that reweights the training loss to reduce correlations between class labels and nuisance attributes. It further applies a penalty based on the *Hilbert-Schmidt Independence Criterion (HSIC)* to explicitly remove nuisance information from learned features, enhancing semantic OOD detection under shared nuisance conditions. **NsED** [26] decomposes images into semantic (phase) and style (amplitude) components via Discrete Fourier Transform, generating augmented samples by mixing amplitude spectra across images. Training with a robust loss that minimizes worst-case classification error over these augmented samples results in features less sensitive to style variations.

**Post-hoc methods** operate after model training to improve OOD detection. **Projection Regret (PR)** [27] uses partial diffusion and denoising to project inputs onto the ID manifold, measuring semantic novelty by comparing the original image with its projection. A second projection step recursively removes background bias, isolating true semantic differences and enabling object-level OOD detection focused on meaningful changes rather than background similarity. **CoVer** [28] enhances detection by evaluating model confidence across multiple corrupted versions of the same input (e.g., fog, blur, contrast shifts). Instead of relying on a single prediction, it averages confidence scores over these variants, improving robustness.

Finally, several **zero-shot OOD detection** approaches, while not explicitly targeting spurious correlations, report results under limited spurious settings. **Maximum Concept Matching (MCM)** [29] leverages CLIP's vision-language embeddings for zero-shot detection without fine-tuning or extra data. Each ID class is represented by a text prompt encoded as a concept prototype, and test images are scored based on cosine similarity with these prototypes. A softmax scaling sharpens distinctions between ID and OOD samples. Another zero-shot multimodal method [30] improves performance by prompting a large language model to generate rich descriptors for each ID class, filtered by consistency on retrieval tasks to reduce hallucinations. It combines CLIP-based similarity scores between test images, filtered descriptors, and detected object labels to compute a robust matching score. **NegLabel** [31] defines negative labels semantically distant from ID classes and computes softmax-based similarity scores that favor ID labels while penalizing negatives, optionally averaging over label groups to reduce noise. Lastly, **CMA** [32] embeds both ID class labels and neutral, class-agnostic Agents in a shared semantic space, leveraging a triangular similarity relationship to reduce confidence on OOD images while maintaining high confidence on ID data.

Together, these approaches represent diverse strategies to tackle the spurious correlation problem in OOD detection, from data augmentation and loss regularization to post-hoc corrections and zero-shot semantic reasoning.

While existing approaches have made significant strides in mitigating spurious correlations in OOD detection, each comes with certain trade-offs and constraints. Outlier exposure and training-time regularization methods often involve non-trivial training overhead, including access to auxiliary data or retraining backbone models with certain objectives, which limits their practicality in deployment settings. Moreover, some of these methods, such as BackMix [24], primarily address specific types of spurious features (e.g., background), which may not generalize to more complex real-world scenarios involving multiple or less structured spurious cues. Post-hoc methods offer a more efficient alternative by avoiding retraining and operating directly on pre-trained models. However, many still rely on careful tuning of multiple hyperparameters or complex transformation pipelines (e.g., CoVer's corruption sets [28] or PR's iterative projections [27]), making them sensitive to design choices and dataset specifics. Furthermore, several zero-shot approaches utilize vision-language models like CLIP, benefiting from rich semantic priors but introducing dependencies on text prompts and external descriptors, which can be limiting when only visual features are available or when text-image alignment is imperfect.

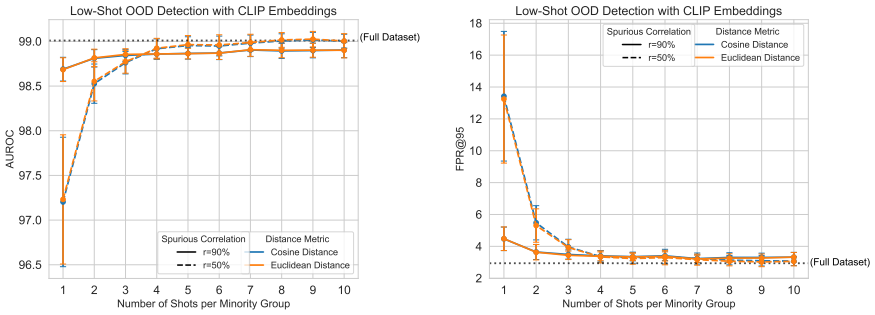
## C Comparison with SP-OOD Methods

Most existing OOD detection methods are not directly comparable to SPROD, as they often rely on different assumptions, such as access to auxiliary OOD data, or necessitate retraining the feature backbone with specific objectives. These design choices contrast with SPROD's post-hoc nature and its objectives of simplicity, efficiency, and broad applicability without model retraining. Nonetheless, to provide a broader context, we include results from a selection of methods that, while not perfectly aligned, have been evaluated under similar SP-OOD conditions. For fairness and consistency, we report the results as published in their original papers, rather than re-implementing them, thereby representing each method according to its best-reported performance. As shown in Table 4, under the most comparable conditions presented, SPROD achieves highly competitive performance, notably without requiring model retraining (when using a ResNet-18 backbone, as applicable to some comparisons) or access to an auxiliary text modality (in the case of methods leveraging CLIP).

To further investigate the sample efficiency of SPROD, we conduct a low-shot experiment using feature embeddings from a frozen CLIP ViT-B/16 vision encoder, evaluated on the Waterbirds dataset. Two experimental settings for spurious correlation are considered: a 50% correlation rate (where majority and minority group sample sizes are equal) and a 90% correlation rate (where majority group samples are nine times more numerous than minority group samples). Within each setting, we vary the number of available samples per minority group and evaluate performance using both Euclidean and cosine distance metrics for score calculation. The results are presented in Figure 7. Initially, with very few samples per minority group, the 90% correlation setting exhibits slightly higher performance, which may be due to the larger initial population of majority group samples aiding prototype stability. However, as the number of samples per minority group increases (e.g., to four samples), the 50% correlation setting surpasses the 90% setting in performance. Notably, the performance in both low-shot variants quickly becomes competitive with the performance achieved

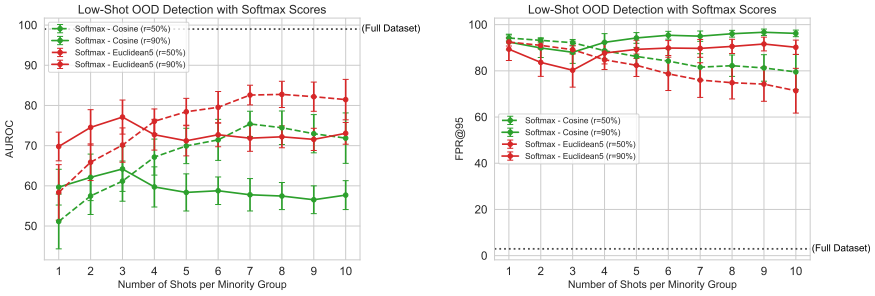
Table 4: A comparative analysis of AUROC and FPR@95 performance metrics for different methods and models evaluated on the Waterbirds dataset. To ensure a fair comparison, we report our results using both the pretrained CLIP and ResNet-18 models, aligning with the settings of the compared methods.

| Method | Backbone | AUROC ↑ | FPR@95 ↓ | Notes |
|---|---|---|---|---|
| Backmix [24] + Energy | WideResNet40-4 | 80.6 | 81.7 | |
| ImOOD [23] | | 83.27 | 57.69 | |
| RW [25] + MD | ResNet-18 | <90 | – | Exact AUROC not clear from plots |
| **SPROD** | | **98.28** | **7.27** | |
| CoVer [28] | | 90.52 | 33.17 | Also reports MCM: 90.31 / 25.66 |
| MCM [29] | | 98.36 | 5.87 | |
| Dai et al. [30] | CLIP-B/16 | 98.62 | 4.56 | |
| Neglabel [31] | | 94.67 | 9.5 | Also reports MCM: 93.30 / 14.45 |
| CMA [32] | | **99.01** | 3.22 | |
| **SPROD** | | **99.01** | **2.94** | |




(a) AUROC (b) FPR@95

Figure 7: Low-shot SP-OOD detection performance of SPROD on the Waterbirds dataset using features from a frozen CLIP ViT-B/16 vision encoder. Performance (AUROC and FPR@95) is shown as a function of the number of samples per minority group for two spurious correlation rates (50% and 90%) and two distance metrics (Euclidean and Cosine). Dashed lines indicate performance with the full training set.




(a) AUROC (b) FPR@95

Figure 8: Impact of softmax normalization on low-shot SP-OOD detection performance of SPROD on the Waterbirds dataset using features from a frozen CLIP ViT-B/16 vision encoder. OOD scores are derived by applying softmax to the negative distances. Performance (AUROC and FPR@95) is shown as a function of the number of samples per minority group for two spurious correlation rates (50% and 90%) and two distance metrics (Euclidean and Cosine). Dashed lines indicate performance with the full training set using softmax-normalized scores.

using the full dataset, underscoring the sample efficiency of the prototypical approach, even with CLIP features.

We extend this low-shot analysis by examining the impact of applying a softmax normalization to the distance-based OOD scores. Figure 8 illustrates the results of this modification. A considerable

performance degradation is observed when softmax normalization is applied to the distance scores, further highlighting the potential sensitivity of softmax-based scoring mechanisms. In this softmax-normalized setting, Euclidean distance appears to yield relatively better performance compared to cosine distance, an observation that differs from the direct distance-based scoring results shown in Figure 7, where both distance metrics performed equally.

# D Dataset Details and Examples

Datasets available for studying spurious correlations are generally limited. In the context of OOD detection, this limitation becomes even more pronounced, as the task requires datasets with SP-OOD samples. As a result, the datasets employed must be thoughtfully considered to ensure meaningful evaluation. In this study, our goal is to address a broader range of spurious features present in data, beyond just background features. A distinctive aspect of our experimental design is the inclusion of three additional settings to explore:

1. **Multi-spurious feature setting**: Scenarios where multiple spurious features (e.g., both background and co-occurring objects) are simultaneously present, increasing the complexity of the detection task. To the best of our knowledge, this is the first time such a setting has been explored in the context of OOD detection.
2. **Multi-class setting**: Scenarios with more than two classes, where inter-class relationships and spurious correlations introduce additional challenges.
3. **Realistic dataset setting**: Beyond existing datasets in the literature, where spurious correlations are often predefined and controlled, we also focus on scenarios leveraging realistic datasets such as AnimalsMetaCoCo and Sp-ImageNet100 [36]. These datasets include diverse spurious correlations and more closely mimic real-world data distributions.

These aspects are relatively underexplored in the context of SP-OOD detection. By incorporating these settings, we aim to demonstrate the effectiveness of our proposed approach.

To evaluate the proposed approach, we used the following datasets:

- **Waterbirds** [17]: This synthetic dataset is generated by combining the CUB [74] (bird classes) and the Places [69] (background scenes) datasets for a binary classification task, with labels $y \in \{\text{waterbird}, \text{landbird}\}$. Spurious correlations are introduced between the background $e \in \{\text{water}, \text{land}\}$ and the label. The dataset consists of four groups, as depicted in Figure 9, with the minority and majority groups highlighted by red and green borders, respectively. Two different correlation values, $r \in \{0.5, 0.9\}$, are employed, where $r$ denotes the probability that the environment $e$ aligns with the label $y$. Specifically, we have:

  $$r = P(e = \text{water} \mid y = \text{waterbird}) = P(e = \text{land} \mid y = \text{landbird}).$$

  The distribution of samples within each group and class is provided in Table 5. For SP-OOD, we select samples from the Places dataset [69], following the previous works [14, 25].
- **CelebA** [34]: This dataset is used for a binary classification task with labels $y \in \{\text{blond hair}, \text{non-blond hair}\}$, where spurious correlations with gender $\in \{\text{male}, \text{female}\}$ are present. It is a real-world dataset, making it suitable for evaluating models in realistic settings. In the dataset, most females have blond hair, and most males have non-blond hair, forming the majority groups, as shown in Figure 10 with colored borders. The minority groups are the opposite combinations. The distribution of groups across classes with varying correlation levels is presented in Table 6. In this setting, SP-OOD samples are those with no hair, but still exhibiting the spurious gender-related features. For our SP-OOD samples, we used bald males, who lack core features but retain spurious (gender) cues.
- **UrbanCars** [35]: This dataset, introduced in [35], is synthetically generated by combining the Stanford Cars dataset [75] (which includes both urban and country cars) with co-occurring objects from either urban or country environments, sourced from the LVIS dataset [76]. The co-occurring objects are positioned to the right of the cars, and both the car and the object are placed onto background images from the Places dataset [69], representing either urban or country scenes. The dataset is considered particularly challenging due to the presence of two spurious features. By varying the combinations of cars, co-occurring

objects, and backgrounds, the dataset is divided into eight distinct groups, as illustrated in Figure 11. Notably, the "all-country" and "all-urban" groups dominate the dataset, while the other groups are underrepresented, as highlighted by the border colors. The exact number of samples in each group of our generated dataset, which exhibits a 95% correlation, is detailed in Table 7, with two groups being especially underrepresented. For the SP-OOD analysis, we sample combinations of backgrounds and co-occurring objects, called the BG & CoObj setting, as well as backgrounds alone, referred to as the BG setting. In the results section, we reported the more challenging scenario(BG & CoObj) as it presented greater difficulty, as expected.

- **AnimalsMetaCoCo**: To create a multi-class, multi-valued spurious attribute setting that reflects realistic and challenging scenarios, we selected a subset of 26 animal classes from the MetaCoCo dataset [71]. The samples were first cleaned through relabeling, removal of duplicates, and deletion of irrelevant images. Subsequently, we defined subattributes to create a total of 8 major concepts for modeling attribute imbalance and spurious features, as illustrated in the first Figure of the paper. We introduce this new dataset as **AnimalsMetaCoCo**, a refined subset tailored for multi-class, multi-valued spurious feature scenarios. Details of the classes, attributes, and group sample distributions are provided in Table 8. This dataset represents a more realistic scenario, where each class is associated with at least one spurious feature similar to those present in ID data. While the strength of spurious correlations may not be as pronounced as in the other three datasets, AnimalsMetaCoCo introduces several unique and challenging aspects. Specifically, its spurious attributes can take on multiple distinct values, each exhibiting varying degrees of imbalance across classes. Moreover, the class distributions themselves are inherently imbalanced. To construct SP-OOD scenarios, we adopt a leave-2-out strategy: in each round, two classes are treated as SP-OOD, characterized by semantic shifts while retaining at least one spurious attribute shared with ID data, and the remaining classes serve as ID. This setup significantly increases the difficulty of the detection task due to overlapping spurious patterns across environments.
- **Sp-ImageNet100**: We evaluate models on the **Spurious ImageNet (SpI)** dataset introduced in [36]. This dataset contains real-world images (from OpenImages [77] and Flickr) that include only spurious features, such as bird feeders or graffiti, without the actual class object. These images are consistently misclassified as specific ImageNet classes, revealing harmful spurious correlations.

  SpI focuses on 100 ImageNet classes (we call Sp-ImageNet100) where such correlations are prevalent. The authors distinguish between two types of harmful spurious features:

  - **Spurious Class Extension:** A spurious feature alone causes a class prediction (e.g., bird feeder $\rightarrow$ hummingbird).
  - **Spurious Shared Feature:** A feature shared across classes biases prediction toward one due to imbalance (e.g., water jet $\rightarrow$ fireboat over fountain).

  We note that most spurious features in SpI are of the class extension type. This is less aligned with our goal of identifying underrepresented groups based on **shared spurious features** and misclassification signals [20], and without group supervision.

  Nevertheless, we include SpI as a challenging, naturally occurring OOD benchmark for evaluating model robustness to spurious correlations.

For NSP-OOD, we utilized the SVHN [78], LSUN [79], and iSUN [80] datasets, which are commonly employed in prior works [14, 25]. These datasets are used consistently as NSP-OOD for all the mentioned ID datasets.

## E Performance on Non-Spurious OOD Datasets

We evaluate the performance of post-hoc methods in the non-spurious OOD (NSP-OOD) detection setting. For this evaluation, the OOD samples are drawn from a collection of standard benchmark datasets: SVHN [78], LSUN (resized)[79], and iSUN[80]. The ID context for each evaluation varies across the five datasets used in our primary spurious correlation experiments: Waterbirds (WB), CelebA (CA), UrbanCars (UC), Animals MetaCoCo (AMC), and Spurious ImageNet (SpI). Specifically, for each of these five ID contexts, the pretrained ResNet-50 backbone is employed, and


Figure 9: Representative examples from the Waterbirds [17] dataset. The dataset consists of four groups: *(Waterbird, Water)*, *(Waterbird, Land)*, *(Landbird, Water)*, and *(Landbird, Land)*. Minority groups, indicated with red borders, are underrepresented, while majority groups, indicated with green borders, are more prevalent. Spurious OOD samples include only background features (land or water) without core bird-related features.

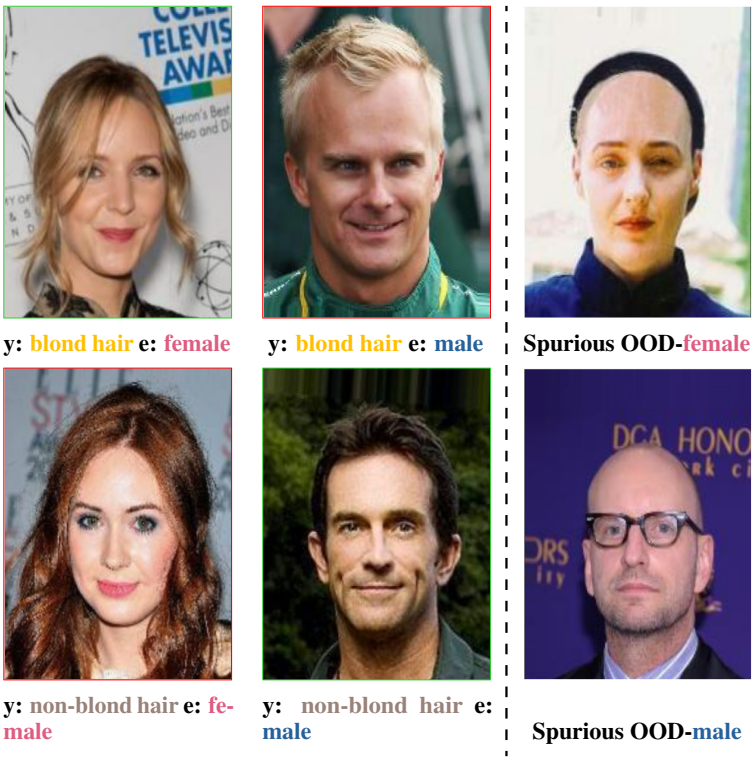



Figure 10: Representative examples from the CelebA [34] dataset, which is divided into four groups: *(Blond hair, Female)*, *(Blond hair, Male)*, *(Non-blond, Female)*, and *(Non-blond, Male)*. Minority groups, marked with red borders, are underrepresented, while majority groups, highlighted with green borders, are more prevalent. Spurious OOD samples contain only spurious features (gender) without core hair color attributes.

Table 5: Group-wise distribution of the **Waterbirds** [17] training set across land and water attributes at varying correlation levels. The distribution reflects the degree of alignment between bird classes (landbird or waterbird) and their respective backgrounds, with higher correlation level indicating a stronger dependence between the bird label and the background.

| Correlation | SP-Feature | Landbird | Waterbird | Total (Row) |
|---|---|---|---|---|
| 50% | Land | 544 | 1853 | **2397** |
| | Water | 545 | 1853 | **2398** |
| | **Total (Col)** | **1089** | **3706** | **4795** |
| 90% | Land | 997 | 369 | **1366** |
| | Water | 111 | 3318 | **3429** |
| | **Total (Col)** | **1108** | **3687** | **4795** |

Table 6: Distribution of the **CelebA** [34] training set across male and female attributes at varying correlation levels. The correlation levels indicate the strength of the spurious relationship between gender and the presence of blond hair, with higher correlations reflecting a stronger association between these attributes in the dataset.

| Correlation | SP-Feature | Blond | Non-blond | Total (Row) |
|---|---|---|---|---|
| 50% | Male | 1387 | 1387 | **2774** |
| | Female | 1387 | 1387 | **2774** |
| | **Total (Col)** | **2774** | **2774** | **5548** |
| 90% | Male | 296 | 2468 | **2764** |
| | Female | 2474 | 310 | **2784** |
| | **Total (Col)** | **2770** | **2778** | **5548** |

Table 7: Distribution of **UrbanCars** [35] training samples with 95% correlation across groups, categorized by background and co-occurring object features within each class. This dataset features six minority groups out of eight possible combinations, highlighting its challenging nature due to the underrepresentation of most groups.

| SP-Features | Country Car | Urban Car | Total (Row) |
|---|---|---|---|
| Country BG, Country CoObj | 3606 | 10 | **3616** |
| Country BG, Urban CoObj | 190 | 190 | **380** |
| Urban BG, Country CoObj | 190 | 189 | **379** |
| Urban BG, Urban CoObj | 10 | 3605 | **3615** |
| **Total (Col)** | **3996** | **3994** | **7990** |

the post-hoc OOD detection methods are subsequently set up using the features derived from this backbone.

The performance metrics reported in Tables 9 and 10 are the average performance when distinguishing each ID dataset from the NSP-OOD samples, averaged across SVHN, LSUN, and iSUN. This setup assesses the general OOD detection capability of the post-hoc methods when the primary challenge is not spurious correlations shared between ID and OOD, but the ID datasets still contain inherent biases. As observed in the tables, distance-based methods such as SPROD and KNN consistently achieve near-perfect scores across all metrics (AUROC, FPR@95, AUPR-In, and AUPR-Out) and ID contexts. Other methods like VIM and MDS also show strong performance.

## F Ablation Study on SPROD Stages

To understand the contribution of each stage in our proposed SPROD method, we conduct an ablation study. We evaluate the performance of:

- Stage 1: Initial Prototype Construction

Table 8: Group-wise distribution of 26 selected animal classes from the MetaCoCo [71] dataset called **Animals MetaCoCo** across various environments.

| Class \Attribute | at home | autumn | dim | grass | in cage | on snow | rock | water | Total |
|---|---|---|---|---|---|---|---|---|---|
| bear | 0 | 109 | 76 | 219 | 0 | 98 | 98 | 362 | 962 |
| cat | 284 | 159 | 150 | 439 | 95 | 102 | 121 | 386 | 1736 |
| cow | 57 | 137 | 119 | 693 | 0 | 122 | 38 | 221 | 1387 |
| crab | 0 | 0 | 43 | 73 | 0 | 0 | 153 | 102 | 371 |
| dog | 106 | 221 | 174 | 561 | 111 | 208 | 145 | 440 | 1966 |
| dolphin | 0 | 0 | 113 | 0 | 0 | 0 | 15 | 425 | 553 |
| elephant | 108 | 98 | 247 | 434 | 0 | 59 | 38 | 375 | 1359 |
| fox | 0 | 195 | 62 | 367 | 66 | 160 | 136 | 145 | 1131 |
| frog | 0 | 232 | 3 | 530 | 0 | 0 | 322 | 342 | 1429 |
| giraffe | 0 | 479 | 150 | 397 | 0 | 0 | 78 | 227 | 1331 |
| goose | 0 | 105 | 135 | 378 | 54 | 0 | 83 | 263 | 1018 |
| horse | 57 | 366 | 128 | 672 | 0 | 129 | 59 | 457 | 1868 |
| kangaroo | 0 | 84 | 190 | 214 | 0 | 55 | 61 | 97 | 701 |
| lion | 0 | 560 | 58 | 537 | 36 | 79 | 275 | 171 | 1716 |
| lizard | 0 | 130 | 42 | 303 | 35 | 0 | 302 | 242 | 1054 |
| monkey | 0 | 183 | 84 | 592 | 76 | 100 | 463 | 255 | 1753 |
| ostrich | 0 | 164 | 125 | 235 | 159 | 76 | 73 | 144 | 976 |
| owl | 0 | 141 | 147 | 131 | 36 | 92 | 78 | 87 | 712 |
| rabbit | 0 | 147 | 31 | 637 | 105 | 134 | 91 | 73 | 1218 |
| rat | 110 | 0 | 0 | 123 | 52 | 41 | 0 | 66 | 392 |
| seal | 0 | 57 | 31 | 158 | 19 | 266 | 240 | 547 | 1318 |
| sheep | 0 | 626 | 30 | 865 | 0 | 99 | 207 | 237 | 2064 |
| squirrel | 0 | 212 | 32 | 418 | 0 | 132 | 188 | 118 | 1100 |
| tiger | 0 | 212 | 14 | 435 | 66 | 176 | 236 | 323 | 1462 |
| tortoise | 0 | 129 | 140 | 284 | 0 | 0 | 157 | 234 | 944 |
| wolf | 0 | 192 | 97 | 198 | 90 | 151 | 188 | 188 | 1104 |
| Total | 722 | 4938 | 2421 | 9893 | 1000 | 2279 | 3845 | 6527 | **31625** |

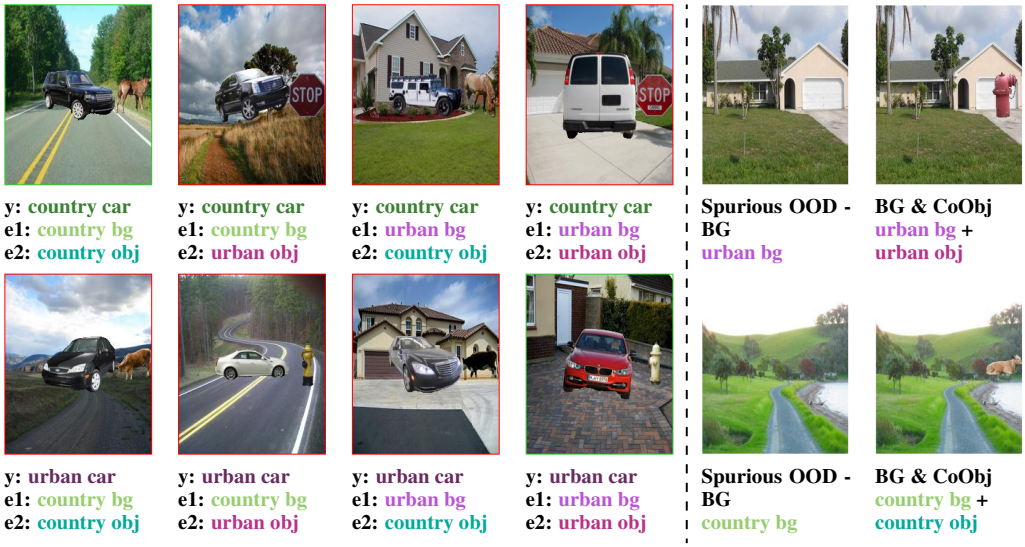



Figure 11: Representative examples from the UrbanCars [35] dataset, which has 8 groups, each pairing country and urban car classes with two spurious features: background and co-occurring object, both of which can be urban or country related. The "all-country" and "all-urban" groups (green borders) dominate the dataset, while the remaining groups are significantly underrepresented (red borders). Spurious OOD samples, containing only spurious features (no cars), consist of combinations of background and co-occurring objects (BG & CoObj) as well as background-only (BG) samples.

Table 9: Average NSP-OOD detection performance (AUROC and FPR@95) using a ResNet-50 backbone. Cell values are performance metrics averaged over SVHN, LSUN, and iSUN as OOD.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | $86.2_{\pm 1.0}$ | $77.5_{\pm 0.9}$ | $52.6_{\pm 4.0}$ | $92.6_{\pm 0.2}$ | $94.9_{\pm 0.3}$ | 80.8 |
| Energy | $85.0_{\pm 6.4}$ | $75.4_{\pm 6.3}$ | $53.2_{\pm 8.0}$ | $96.8_{\pm 0.1}$ | $97.7_{\pm 0.3}$ | 81.6 |
| MLS | $85.7_{\pm 4.2}$ | $76.4_{\pm 4.1}$ | $53.0_{\pm 6.3}$ | $96.2_{\pm 0.1}$ | $96.3_{\pm 0.3}$ | 81.5 |
| KLM | $57.7_{\pm 2.0}$ | $52.5_{\pm 2.1}$ | $47.5_{\pm 1.6}$ | $90.1_{\pm 0.3}$ | $96.1_{\pm 0.5}$ | 68.8 |
| GNorm | $90.7_{\pm 0.6}$ | $70.8_{\pm 0.9}$ | $64.1_{\pm 2.1}$ | $97.3_{\pm 0.1}$ | $99.0_{\pm 0.0}$ | 84.4 |
| ReAct | $86.8_{\pm 5.2}$ | $66.1_{\pm 7.5}$ | $50.0_{\pm 8.7}$ | $96.3_{\pm 0.1}$ | $99.0_{\pm 0.3}$ | 79.6 |
| VIM | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $99.6_{\pm 0.4}$ | $99.9_{\pm 0.0}$ | $99.6_{\pm 0.0}$ | 99.8 |
| MDS | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $99.4_{\pm 0.0}$ | $97.5_{\pm 0.0}$ | 99.4 |
| RMDS | $58.6_{\pm 0.4}$ | $68.6_{\pm 2.9}$ | $48.9_{\pm 0.6}$ | $93.0_{\pm 0.2}$ | $97.1_{\pm 0.1}$ | 73.2 |
| KNN | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $99.8_{\pm 0.0}$ | **100.0** |
| SHE | $96.4_{\pm 0.2}$ | $98.6_{\pm 0.2}$ | $99.1_{\pm 0.1}$ | $94.3_{\pm 0.2}$ | $98.5_{\pm 0.1}$ | 97.4 |
| **SPROD** | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{99.9}_{\pm 0.0}$ | **100.0** |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | $76.5_{\pm 1.6}$ | $85.2_{\pm 2.7}$ | $95.6_{\pm 0.4}$ | $46.1_{\pm 1.5}$ | $34.8_{\pm 3.1}$ | 67.6 |
| Energy | $74.5_{\pm 24.8}$ | $84.4_{\pm 14.5}$ | $90.8_{\pm 9.7}$ | $19.9_{\pm 1.1}$ | $13.4_{\pm 2.2}$ | 56.6 |
| MLS | $76.3_{\pm 15.0}$ | $86.9_{\pm 9.6}$ | $93.8_{\pm 5.0}$ | $26.3_{\pm 1.3}$ | $24.5_{\pm 2.5}$ | 61.6 |
| KLM | $80.5_{\pm 1.4}$ | $84.9_{\pm 2.6}$ | $95.5_{\pm 0.5}$ | $57.8_{\pm 1.4}$ | $25.5_{\pm 6.2}$ | 68.8 |
| GNorm | $66.4_{\pm 2.7}$ | $85.2_{\pm 2.1}$ | $93.2_{\pm 0.6}$ | $13.8_{\pm 0.6}$ | $4.8_{\pm 0.3}$ | 52.7 |
| ReAct | $73.5_{\pm 23.4}$ | $88.0_{\pm 10.7}$ | $91.6_{\pm 6.7}$ | $23.3_{\pm 0.8}$ | $4.3_{\pm 2.0}$ | 56.1 |
| VIM | $\mathbf{0.0}_{\pm 0.0}$ | $\mathbf{0.0}_{\pm 0.0}$ | $1.9_{\pm 1.6}$ | $\mathbf{0.0}_{\pm 0.0}$ | $\mathbf{0.0}_{\pm 0.0}$ | 0.4 |
| MDS | $\mathbf{0.0}_{\pm 0.0}$ | $\mathbf{0.0}_{\pm 0.0}$ | $\mathbf{0.0}_{\pm 0.0}$ | $0.7_{\pm 0.0}$ | $11.9_{\pm 0.4}$ | 2.5 |
| RMDS | $90.5_{\pm 0.3}$ | $89.2_{\pm 1.7}$ | $95.4_{\pm 0.4}$ | $51.3_{\pm 1.6}$ | $17.7_{\pm 0.4}$ | 68.8 |
| KNN | $\mathbf{0.0}_{\pm 0.0}$ | $\mathbf{0.0}_{\pm 0.0}$ | $\mathbf{0.0}_{\pm 0.0}$ | $\mathbf{0.0}_{\pm 0.0}$ | $0.6_{\pm 0.0}$ | 0.1 |
| SHE | $15.9_{\pm 0.6}$ | $7.0_{\pm 0.8}$ | $4.9_{\pm 0.4}$ | $29.8_{\pm 1.0}$ | $7.7_{\pm 0.8}$ | 13.1 |
| **SPROD** | $\mathbf{0.0}_{\pm 0.0}$ | $\mathbf{0.0}_{\pm 0.0}$ | $\mathbf{0.0}_{\pm 0.0}$ | $\mathbf{0.0}_{\pm 0.0}$ | $\mathbf{0.0}_{\pm 0.0}$ | **0.0** |

Table 10: Average NSP-OOD detection performance (AUPR-In and AUPR-Out) using a ResNet-50 backbone. Cell values are performance metrics averaged over SVHN, LSUN, and iSUN as OOD.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | $95.8_{\pm 0.3}$ | $67.7_{\pm 1.4}$ | $43.4_{\pm 3.9}$ | $96.0_{\pm 0.1}$ | $95.6_{\pm 0.2}$ | 79.7 |
| Energy | $95.5_{\pm 1.8}$ | $66.2_{\pm 5.0}$ | $44.2_{\pm 6.2}$ | $98.3_{\pm 0.1}$ | $97.9_{\pm 0.2}$ | 80.4 |
| MLS | $95.7_{\pm 1.3}$ | $66.5_{\pm 4.0}$ | $43.9_{\pm 5.6}$ | $98.0_{\pm 0.1}$ | $96.8_{\pm 0.2}$ | 80.2 |
| KLM | $77.1_{\pm 1.1}$ | $27.8_{\pm 1.0}$ | $34.4_{\pm 1.3}$ | $93.9_{\pm 0.2}$ | $96.6_{\pm 0.4}$ | 66.0 |
| GNorm | $97.2_{\pm 0.2}$ | $49.7_{\pm 1.1}$ | $59.6_{\pm 2.0}$ | $98.4_{\pm 0.1}$ | $99.0_{\pm 0.0}$ | 80.8 |
| ReAct | $96.1_{\pm 1.5}$ | $53.9_{\pm 4.6}$ | $40.6_{\pm 5.9}$ | $98.0_{\pm 0.0}$ | $99.0_{\pm 0.3}$ | 77.5 |
| VIM | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $99.2_{\pm 0.7}$ | $\mathbf{100.0}_{\pm 0.0}$ | $99.7_{\pm 0.0}$ | 99.8 |
| MDS | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $99.7_{\pm 0.0}$ | $98.0_{\pm 0.0}$ | 99.5 |
| RMDS | $78.0_{\pm 0.2}$ | $46.1_{\pm 3.9}$ | $35.8_{\pm 0.5}$ | $96.4_{\pm 0.1}$ | $97.4_{\pm 0.1}$ | 70.7 |
| KNN | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $99.8_{\pm 0.0}$ | **100.0** |
| SHE | $98.7_{\pm 0.1}$ | $96.6_{\pm 0.4}$ | $98.6_{\pm 0.1}$ | $89.6_{\pm 0.5}$ | $98.6_{\pm 0.1}$ | 96.4 |
| **SPROD** | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{99.9}_{\pm 0.0}$ | **100.0** |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | $55.8_{\pm 1.9}$ | $85.9_{\pm 0.7}$ | $65.0_{\pm 2.2}$ | $85.4_{\pm 0.6}$ | $93.9_{\pm 0.5}$ | 77.2 |
| Energy | $56.4_{\pm 18.0}$ | $83.6_{\pm 6.8}$ | $66.3_{\pm 7.9}$ | $92.6_{\pm 0.4}$ | $97.5_{\pm 0.3}$ | 79.3 |
| MLS | $56.3_{\pm 12.1}$ | $84.7_{\pm 4.4}$ | $65.7_{\pm 5.5}$ | $91.3_{\pm 0.4}$ | $95.7_{\pm 0.4}$ | 78.7 |
| KLM | $38.4_{\pm 1.7}$ | $76.9_{\pm 0.9}$ | $62.7_{\pm 0.6}$ | $79.6_{\pm 0.6}$ | $94.9_{\pm 0.6}$ | 70.5 |
| GNorm | $65.1_{\pm 1.7}$ | $83.8_{\pm 0.7}$ | $71.1_{\pm 1.5}$ | $94.2_{\pm 0.2}$ | $98.9_{\pm 0.0}$ | 82.6 |
| ReAct | $57.9_{\pm 15.9}$ | $78.6_{\pm 7.4}$ | $65.0_{\pm 8.0}$ | $91.8_{\pm 0.2}$ | $99.0_{\pm 0.3}$ | 78.5 |
| VIM | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $99.8_{\pm 0.2}$ | $99.9_{\pm 0.0}$ | $99.2_{\pm 0.0}$ | 99.8 |
| MDS | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $98.8_{\pm 0.0}$ | $96.5_{\pm 0.0}$ | 99.1 |
| RMDS | $31.3_{\pm 0.3}$ | $81.8_{\pm 1.8}$ | $64.0_{\pm 0.6}$ | $83.5_{\pm 0.5}$ | $96.3_{\pm 0.1}$ | 71.4 |
| KNN | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $99.8_{\pm 0.0}$ | **100.0** |
| SHE | $91.6_{\pm 0.4}$ | $99.4_{\pm 0.1}$ | $99.5_{\pm 0.0}$ | $97.0_{\pm 0.1}$ | $98.3_{\pm 0.1}$ | 97.2 |
| **SPROD** | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | $99.9_{\pm 0.0}$ | $\mathbf{100.0}_{\pm 0.0}$ | **100.0** |

- Stage 2: Classification-Aware Prototype Calculation
- Stage 3: Group Prototype Refinement

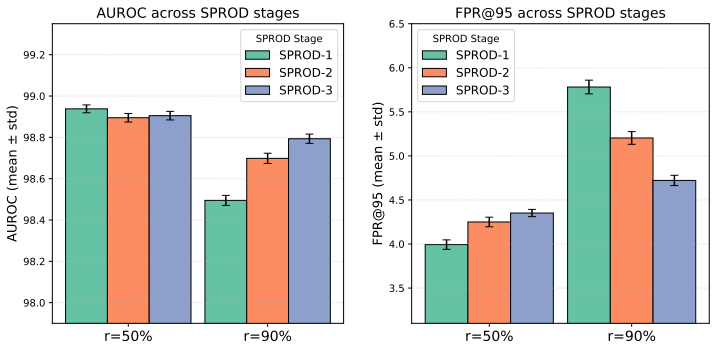
For this analysis, we select the Waterbirds dataset, as its synthetic nature allows for precise control over spurious correlations and clearly exemplifies the SP-OOD challenge by design. We evaluate on two versions of Waterbirds: one with a 50% spurious correlation rate (where spurious features are less effective) and another with a 90% correlation rate (representing a strong spurious bias). Experiments are conducted using features from pretrained ResNet-50 and ResNet-18 backbones, without fine-tuning on Waterbirds, to isolate the effect of the prototype refinement stages.

The results of this ablation study are presented in Figure 12 and Figure 13. As seen, the performance of the simple initial prototypical approach (Stage 1) performs competitively, especially on the 50% correlation setting. This suggests that a basic prototypical method, which computes distances to class means in feature space, is a competitive baseline for OOD detection that has been somewhat overlooked in existing literature. When the spurious correlation rate is increased to 90%, we observe a general reduction in OOD detection performance across all three variants. This is expected, as stronger spurious correlations make it more difficult to distinguish true class features from misleading cues. However, Stage 3 (the full SPROD method) is significantly more robust to the increase of spurious correlation.

# G Backbone Experiments

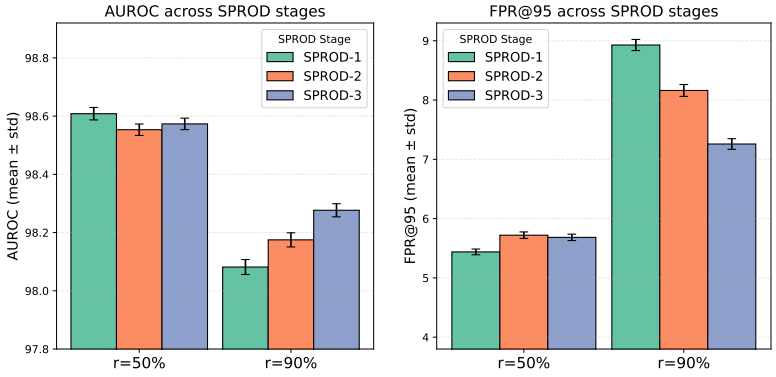
To assess the generality and robustness of SPROD across different neural network architectures, we evaluate its performance using a wide range of feature backbones. This analysis complements the main paper's results, which primarily rely on ResNet-50 [68], and demonstrates that our method remains effective across both convolutional and transformer-based representations.

We include a comprehensive selection of backbones commonly used in the OOD detection literature. This includes all major variants of the ResNet [68] family, ResNet-18 (R18), ResNet-34 (R34), ResNet-50 (R50), and ResNet-101 (R101), owing to their widespread adoption and varying representational capacities. Alongside these, we evaluate several modern transformer-based architectures with diverse embedding sizes and training paradigms, such as the self-supervised DINOv2-S (DINOv2) [81], the standard supervised ViT-S (ViT) [82], and more hierarchical or data-efficient




(a) AUROC (b) FPR@95

Figure 12: Ablation study of SPROD stages on the Waterbirds dataset using ResNet-50 features. Results are shown for two spurious correlation settings: 50% and 90%.




(a) AUROC (b) FPR@95

Figure 13: Ablation study of SPROD stages on the Waterbirds dataset using ResNet-18 features. Results are shown for two spurious correlation settings: 50% and 90%.

designs like Swin-B (Swin) [83], DeiT-B (DeiT) [82], ConvNeXt-B (CvNXt) [84], and BiT-R50x1 (BiT) [85].

We analyze the performance of each method using four complementary metrics (AUROC, FPR@95, AUPR-IN, and AUPR-OUT) to evaluate different aspects of OOD detection performance. Tables 11, 12, 13, and 14 summarize the performance of our post-hoc method across various backbone architectures, averaged over the five SP-OOD datasets studied in this work. Detailed results for each individual backbone across the different datasets are presented in the subsequent tables.

Overall, we observe that SPROD consistently achieves strong performance across all backbone architectures, often with a notable margin. KNN emerges as the second-best method, suggesting that simpler approaches can be counterintuitively effective; however, its performance is highly sensitive to the choice of hyperparameters. MDS is also among the top-performing methods; it is a metric-based approach similar to SPROD but employs a more complex model by estimating class-specific covariance matrices. This added complexity, while potentially beneficial, may increase the risk of overfitting, especially in OOD settings. In addition, output-based methods show notable drops in performance on several benchmarks, highlighting their susceptibility to distributional shifts. Notably, SPROD is the only method that maintains stable performance across all evaluated settings, without experiencing major degradation under any backbone or dataset configuration. Among the evaluated backbones, ConvNeXt-B and Swin-B stand out as frozen feature extractors with superior performance for post-hoc OOD detection.

Table 11: AUROC performance of all evaluated methods across various backbone architectures. Results are averaged over five SP-OOD datasets.

| Method | R18 | R34 | R50 | R101 | DINOv2 | ViT | Swin | DeiT | CvNXt | BiT | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MSP | 61.1 | 61.9 | 61.9 | 63.3 | 70.2 | 65.4 | 70.8 | 69.7 | 72.0 | 65.0 | 66.1 |
| Energy | 61.5 | 61.6 | 61.3 | 62.7 | 70.5 | 65.8 | 70.4 | 70.0 | 71.5 | 64.7 | 66.0 |
| MLS | 61.7 | 61.9 | 61.6 | 63.1 | 70.5 | 65.8 | 70.7 | 70.1 | 71.3 | 64.8 | 66.2 |
| KLM | 56.6 | 58.8 | 60.7 | 61.5 | 70.2 | 64.5 | 68.6 | 62.4 | 69.3 | 60.6 | 63.3 |
| GNorm | 64.3 | 64.2 | 64.7 | 66.4 | 69.9 | 66.4 | 56.7 | 62.1 | 67.8 | 54.2 | 63.7 |
| ReAct | 62.4 | 62.1 | 64.7 | 64.1 | 70.0 | 67.3 | 72.1 | 69.8 | 70.7 | 60.5 | 66.4 |
| VIM | 66.1 | 68.0 | 69.3 | 72.1 | 78.2 | 73.5 | 77.6 | 72.9 | 80.0 | 77.8 | 73.5 |
| MDS | 65.6 | 70.5 | 72.2 | 74.2 | 83.1 | 82.9 | 80.0 | 71.3 | 84.8 | 84.2 | 76.9 |
| RMDS | 54.8 | 57.2 | 60.2 | 58.3 | 63.3 | 61.1 | 69.6 | 65.5 | 69.5 | 66.5 | 62.6 |
| KNN | 77.7 | 78.6 | 80.3 | 79.9 | 79.7 | 80.3 | 84.3 | 78.9 | 86.1 | 81.6 | 80.7 |
| SHE | 69.4 | 72.4 | 68.4 | 73.5 | 83.0 | 79.8 | 84.9 | 81.8 | 84.6 | 73.5 | 77.1 |
| **SPROD** | **83.0** | **83.2** | **85.1** | **85.9** | **87.2** | **85.1** | **90.1** | **84.6** | **89.8** | **87.1** | **86.1** |

Table 12: FPR@95 performance of all evaluated methods across various backbone architectures. Results are averaged over five SP-OOD datasets.

| Method | R18 | R34 | R50 | R101 | DINOv2 | ViT | Swin | DeiT | CvNXt | BiT | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MSP | 90.8 | 88.9 | 88.4 | 86.5 | 81.1 | 86.4 | 71.7 | 80.4 | 67.8 | 84.2 | 82.6 |
| Energy | 88.5 | 86.9 | 88.9 | 87.5 | 77.9 | 84.9 | 73.7 | 78.6 | 68.2 | 82.8 | 81.8 |
| MLS | 89.6 | 87.4 | 88.5 | 86.8 | 78.3 | 84.9 | 71.5 | 78.2 | 67.8 | 82.7 | 81.6 |
| KLM | 91.3 | 90.3 | 88.3 | 87.4 | 80.6 | 85.7 | 72.6 | 80.5 | 66.8 | 83.5 | 82.7 |
| GNorm | 86.1 | 84.0 | 83.8 | 82.9 | 80.2 | 85.3 | 79.2 | 82.3 | 70.7 | 87.7 | 82.2 |
| ReAct | 85.6 | 84.4 | 84.0 | 83.5 | 78.3 | 86.8 | 72.6 | 78.3 | 69.4 | 86.1 | 80.9 |
| VIM | 83.0 | 81.9 | 78.5 | 74.3 | 56.7 | 75.0 | 68.3 | 82.6 | 47.4 | 61.8 | 71.0 |
| MDS | 82.0 | 80.8 | 73.5 | 67.7 | 51.4 | 61.9 | 70.7 | 82.6 | 50.4 | 48.4 | 66.9 |
| RMDS | 93.2 | 92.2 | 91.6 | 89.4 | 81.0 | 83.8 | 72.9 | 85.2 | 67.7 | 82.2 | 83.9 |
| KNN | 64.7 | 62.4 | 58.4 | 57.3 | 51.8 | 66.7 | 53.4 | 71.9 | 46.9 | 51.0 | 58.5 |
| SHE | 68.2 | 65.7 | 70.5 | 67.0 | 54.2 | 63.1 | 46.4 | 66.9 | 50.5 | 74.0 | 62.6 |
| **SPROD** | **53.1** | **54.6** | **49.0** | **46.6** | **43.0** | **51.1** | **35.3** | **55.1** | **35.2** | **42.9** | **46.6** |

# H Analysis of Mixture of Prototypes

To evaluate the role of prototype augmentation and refinement, here we analyze two new variants of SPROD, comparing them against our standard method (referred to as SPROD-Default in this section). The first is SPROD-KMeans, a clustering baseline where embeddings of training samples within each class are clustered using K-Means. The number of centroids per class is set to match the number of group prototypes that SPROD-Default would derive. This tests whether the benefits of SPROD-Default stem solely from using multiple prototypes per class or from its proposed classification-driven refinement strategy. Recent works, such as Prototypical Learning with a Mixture of Prototypes (PALM) [56], demonstrate the importance of multiple prototypes to capture intra-class diversity. SPROD-Kmeans serves as a simpler post-hoc method based on the K-Means algorithm in this context.

We also introduce SPROD-Converged, which iteratively refines the group prototypes derived from SPROD-Default's Stage 2 by repeatedly applying the Stage 3 reassignment and recalculation steps until centroid convergence, similar to the K-Means algorithm. The centroid initialization mechanism distinguishes SPROD-Converged from SPROD-Kmeans, where K-Means clustering is typically initialized using a method like K-Means++ directly on the raw class samples.

As shown in Tables 35 and 36, SPROD-Default generally outperforms the two other variants when the amount of spurious correlation in the ID data is high (e.g., Waterbirds, CelebA, and UrbanCars); on other datasets, their performances are competitive. Proto-KMeans is slightly less effective in separating SP-OOD samples, likely due to its reliance on purely geometric clustering, whereas SPROD with classification-aware refinement slightly enhances the robustness when the spurious correlation is high.

Table 13: AUPR-IN performance of all evaluated methods across various backbone architectures. Results are averaged over five SP-OOD datasets.

| Method | R18 | R34 | R50 | R101 | DINOv2 | ViT | Swin | DeiT | CvNXt | BiT | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MSP | 48.4 | 49.3 | 49.8 | 50.9 | 59.3 | 54.4 | 61.1 | 59.0 | 64.1 | 51.3 | 54.8 |
| Energy | 47.9 | 48.3 | 48.7 | 50.1 | 59.9 | 54.8 | 59.3 | 58.7 | 63.1 | 50.4 | 54.1 |
| MLS | 48.4 | 48.6 | 49.3 | 50.6 | 59.2 | 54.8 | 60.1 | 59.0 | 63.0 | 50.6 | 54.4 |
| KLM | 41.0 | 43.8 | 45.2 | 45.2 | 56.1 | 49.3 | 52.4 | 48.4 | 52.8 | 45.5 | 48.0 |
| GNorm | 50.3 | 50.2 | 51.0 | 51.3 | 58.5 | 55.5 | 38.8 | 42.9 | 51.5 | 40.5 | 49.0 |
| ReAct | 49.9 | 50.7 | 52.7 | 53.1 | 59.6 | 56.1 | 63.0 | 59.3 | 62.6 | 47.8 | 55.5 |
| VIM | 51.6 | 53.9 | 55.4 | 59.7 | 66.7 | 62.1 | 68.9 | 63.1 | 72.5 | 64.4 | 61.8 |
| MDS | 52.2 | 60.2 | 61.6 | 64.1 | 71.8 | 74.5 | 72.0 | 60.8 | 77.0 | 75.3 | 67.0 |
| RMDS | 42.2 | 45.3 | 45.1 | 45.5 | 50.2 | 49.5 | 59.0 | 53.7 | 59.2 | 48.8 | 49.9 |
| KNN | 63.6 | 66.0 | 68.0 | 67.7 | 64.6 | 70.8 | 74.6 | 69.2 | 76.5 | 69.8 | 69.1 |
| SHE | 52.8 | 57.2 | 52.5 | 56.6 | 75.1 | 71.5 | 75.0 | 73.5 | 80.1 | 61.2 | 65.5 |
| **SPROD** | **72.7** | **74.0** | **76.5** | **77.1** | **78.9** | **77.1** | **84.1** | **76.5** | **84.3** | **78.3** | **78.0** |

Table 14: AUPR-OUT performance of all evaluated methods across various backbone architectures. Results are averaged over five SP-OOD datasets.

| Method | R18 | R34 | R50 | R101 | DINOv2 | ViT | Swin | DeiT | CvNXt | BiT | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| MSP | 72.2 | 73.0 | 73.0 | 74.3 | 79.9 | 75.3 | 81.3 | 78.9 | 81.9 | 76.3 | 76.6 |
| Energy | 73.1 | 73.6 | 73.0 | 74.0 | 80.4 | 76.1 | 81.5 | 80.0 | 81.9 | 76.6 | 77.0 |
| MLS | 72.9 | 73.5 | 73.1 | 74.2 | 80.4 | 76.0 | 81.6 | 79.8 | 81.8 | 76.6 | 77.0 |
| KLM | 70.8 | 71.5 | 73.0 | 73.8 | 79.3 | 74.5 | 79.6 | 75.4 | 80.9 | 74.0 | 75.3 |
| GNorm | 74.3 | 74.8 | 75.0 | 76.1 | 79.5 | 76.0 | 76.1 | 76.6 | 79.2 | 71.8 | 75.9 |
| ReAct | 73.8 | 74.1 | 75.2 | 75.4 | 80.2 | 76.7 | 82.1 | 80.0 | 81.4 | 73.5 | 77.2 |
| VIM | 77.8 | 79.0 | 80.1 | 81.8 | 86.6 | 82.5 | 84.8 | 80.2 | 89.2 | 86.1 | 82.8 |
| MDS | 77.8 | 79.7 | 81.3 | 82.6 | 88.7 | 87.9 | 83.9 | 78.3 | 90.3 | 88.5 | 83.9 |
| RMDS | 68.9 | 70.4 | 71.8 | 72.4 | 76.4 | 74.0 | 79.8 | 75.7 | 80.7 | 74.9 | 74.5 |
| KNN | 85.6 | 85.9 | 87.4 | 87.7 | 89.4 | 86.9 | 90.7 | 85.8 | 91.1 | 89.5 | 88.0 |
| SHE | 80.1 | 82.6 | 79.1 | 82.6 | 88.3 | 86.4 | 90.7 | 84.6 | 87.6 | 82.4 | 84.4 |
| **SPROD** | **88.8** | **88.7** | **90.1** | **90.7** | **92.9** | **90.3** | **93.9** | **90.7** | **93.8** | **92.7** | **91.3** |

# I Reproducibility and Resources

The data (cleaned AnimalsMetaCoCo dataset) and code for our approach are available at the following GitHub repository: `https://github.com/ReihanehZohrabi/SPROD`. In our benchmarking, we utilized components of the OpenOOD v1.5 [86, 37, 87] framework to obtain results for previously proposed OOD detection methods.

Our experiments were designed to be post-hoc and computationally efficient. All experiments were conducted on a single GeForce RTX 3090 Ti GPU, demonstrating the method's practicality in terms of resource requirements.

Table 15: AUROC and FPR@95 performance of all methods using ResNet-18 as the feature backbone.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 68.1$_{\pm 0.7}$ | 47.6$_{\pm 2.9}$ | 40.2$_{\pm 0.5}$ | 72.4$_{\pm 0.6}$ | 77.3$_{\pm 0.2}$ | 61.1 |
| Energy | 69.4$_{\pm 4.3}$ | 48.5$_{\pm 4.2}$ | 43.2$_{\pm 5.0}$ | 71.0$_{\pm 0.8}$ | 75.6$_{\pm 0.3}$ | 61.5 |
| MLS | 69.2$_{\pm 3.6}$ | 48.1$_{\pm 4.0}$ | 42.7$_{\pm 3.8}$ | 71.6$_{\pm 0.8}$ | 76.8$_{\pm 0.3}$ | 61.7 |
| KLM | 43.9$_{\pm 0.5}$ | 41.1$_{\pm 1.4}$ | 57.7$_{\pm 0.4}$ | 68.2$_{\pm 0.5}$ | 72.3$_{\pm 0.4}$ | 56.6 |
| GNorm | 81.9$_{\pm 0.3}$ | 42.0$_{\pm 2.4}$ | 48.3$_{\pm 0.4}$ | 67.5$_{\pm 0.6}$ | 82.0$_{\pm 0.2}$ | 64.3 |
| ReAct | 75.9$_{\pm 4.1}$ | 40.8$_{\pm 3.7}$ | 44.0$_{\pm 5.5}$ | 71.3$_{\pm 0.8}$ | 79.8$_{\pm 0.4}$ | 62.4 |
| VIM | 77.2$_{\pm 3.0}$ | 52.7$_{\pm 3.8}$ | 63.9$_{\pm 3.8}$ | 71.8$_{\pm 0.8}$ | 64.8$_{\pm 0.2}$ | 66.1 |
| MDS | 74.4$_{\pm 0.1}$ | 58.9$_{\pm 0.5}$ | 89.5$_{\pm 0.1}$ | 60.1$_{\pm 0.8}$ | 44.9$_{\pm 0.1}$ | 65.6 |
| RMDS | 46.3$_{\pm 0.1}$ | 37.1$_{\pm 0.7}$ | 44.7$_{\pm 0.2}$ | 74.8$_{\pm 0.5}$ | 71.0$_{\pm 0.1}$ | 54.8 |
| KNN | 93.9$_{\pm 0.0}$ | 58.8$_{\pm 0.4}$ | 92.1$_{\pm 0.8}$ | 73.3$_{\pm 0.6}$ | 70.3$_{\pm 0.1}$ | 77.7 |
| SHE | 91.0$_{\pm 0.1}$ | 47.4$_{\pm 0.2}$ | 74.2$_{\pm 0.1}$ | 52.3$_{\pm 0.9}$ | **82.3**$_{\pm 0.2}$ | 69.4 |
| **SPROD** | **98.3**$_{\pm 0.0}$ | **63.6**$_{\pm 1.0}$ | **97.3**$_{\pm 0.0}$ | **75.5**$_{\pm 0.5}$ | 80.1$_{\pm 0.0}$ | **83.0** |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 89.7$_{\pm 0.6}$ | 98.0$_{\pm 0.5}$ | 95.5$_{\pm 0.6}$ | 89.2$_{\pm 0.4}$ | 81.7$_{\pm 0.8}$ | 90.8 |
| Energy | 85.7$_{\pm 8.0}$ | 96.9$_{\pm 1.0}$ | 91.3$_{\pm 5.1}$ | 90.8$_{\pm 0.4}$ | 77.9$_{\pm 1.3}$ | 88.5 |
| MLS | 86.7$_{\pm 6.2}$ | 97.8$_{\pm 0.6}$ | 93.0$_{\pm 3.6}$ | 90.3$_{\pm 0.4}$ | 80.3$_{\pm 0.6}$ | 89.6 |
| KLM | 91.2$_{\pm 0.5}$ | 98.0$_{\pm 0.5}$ | 95.5$_{\pm 0.6}$ | 85.8$_{\pm 0.4}$ | 86.0$_{\pm 1.6}$ | 91.3 |
| GNorm | 85.9$_{\pm 0.8}$ | 98.1$_{\pm 0.4}$ | 95.1$_{\pm 0.5}$ | 91.4$_{\pm 0.4}$ | 60.1$_{\pm 1.0}$ | 86.1 |
| ReAct | 82.0$_{\pm 9.1}$ | 98.0$_{\pm 0.8}$ | 92.1$_{\pm 4.5}$ | 90.9$_{\pm 0.5}$ | 64.8$_{\pm 1.5}$ | 85.6 |
| VIM | 69.0$_{\pm 5.6}$ | 95.1$_{\pm 1.0}$ | 71.4$_{\pm 5.7}$ | 89.8$_{\pm 0.5}$ | 89.8$_{\pm 0.5}$ | 83.0 |
| MDS | 69.1$_{\pm 0.1}$ | 94.5$_{\pm 0.5}$ | 58.5$_{\pm 0.5}$ | 92.8$_{\pm 0.3}$ | 94.9$_{\pm 0.0}$ | 82.0 |
| RMDS | 95.0$_{\pm 0.0}$ | 99.0$_{\pm 0.3}$ | 97.2$_{\pm 0.0}$ | 85.8$_{\pm 0.7}$ | 89.2$_{\pm 0.2}$ | 93.2 |
| KNN | 25.1$_{\pm 0.1}$ | 93.7$_{\pm 1.1}$ | 41.7$_{\pm 2.3}$ | 85.7$_{\pm 0.7}$ | 77.2$_{\pm 0.2}$ | 64.7 |
| SHE | 27.7$_{\pm 0.3}$ | 93.9$_{\pm 0.8}$ | 68.4$_{\pm 0.4}$ | 94.3$_{\pm 0.4}$ | **56.6**$_{\pm 0.5}$ | 68.2 |
| **SPROD** | **7.3**$_{\pm 0.1}$ | **91.8**$_{\pm 1.1}$ | **18.8**$_{\pm 0.4}$ | **79.5**$_{\pm 0.6}$ | 67.9$_{\pm 0.2}$ | **53.1** |

Table 16: AUROC and FPR@95 performance of all methods using ResNet-34 as the feature backbone.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 66.0$_{\pm 0.7}$ | 41.9$_{\pm 1.5}$ | 42.7$_{\pm 0.6}$ | 77.5$_{\pm 0.4}$ | 81.6$_{\pm 0.2}$ | 61.9 |
| Energy | 64.9$_{\pm 6.8}$ | 43.2$_{\pm 4.8}$ | 42.5$_{\pm 4.4}$ | 78.2$_{\pm 0.6}$ | 79.4$_{\pm 0.2}$ | 61.6 |
| MLS | 65.0$_{\pm 5.7}$ | 42.6$_{\pm 4.1}$ | 42.5$_{\pm 3.4}$ | 78.4$_{\pm 0.6}$ | 80.8$_{\pm 0.2}$ | 61.9 |
| KLM | 42.2$_{\pm 0.8}$ | 45.0$_{\pm 1.3}$ | 56.2$_{\pm 0.5}$ | 72.2$_{\pm 0.5}$ | 78.2$_{\pm 0.8}$ | 58.8 |
| GNorm | 77.9$_{\pm 0.4}$ | 35.4$_{\pm 0.9}$ | 50.0$_{\pm 0.5}$ | 73.7$_{\pm 0.4}$ | **84.1**$_{\pm 0.1}$ | 64.2 |
| ReAct | 72.3$_{\pm 5.9}$ | 34.4$_{\pm 5.9}$ | 43.5$_{\pm 5.4}$ | 77.9$_{\pm 0.6}$ | 82.4$_{\pm 0.4}$ | 62.1 |
| VIM | 80.8$_{\pm 4.3}$ | 50.3$_{\pm 3.2}$ | 62.4$_{\pm 3.3}$ | 75.5$_{\pm 0.7}$ | 70.9$_{\pm 0.2}$ | 68.0 |
| MDS | 85.7$_{\pm 0.0}$ | 60.8$_{\pm 0.8}$ | 89.8$_{\pm 0.1}$ | 61.4$_{\pm 0.8}$ | 54.8$_{\pm 0.1}$ | 70.5 |
| RMDS | 52.4$_{\pm 0.1}$ | 33.3$_{\pm 1.2}$ | 44.1$_{\pm 0.2}$ | **79.3**$_{\pm 0.6}$ | 77.0$_{\pm 0.1}$ | 57.2 |
| KNN | 97.8$_{\pm 0.0}$ | 56.4$_{\pm 0.7}$ | 88.8$_{\pm 0.2}$ | 75.9$_{\pm 0.7}$ | 73.9$_{\pm 0.1}$ | 78.6 |
| SHE | 87.9$_{\pm 0.2}$ | 40.8$_{\pm 0.3}$ | 81.2$_{\pm 0.2}$ | 68.5$_{\pm 0.2}$ | 83.4$_{\pm 0.2}$ | 72.4 |
| **SPROD** | **98.2**$_{\pm 0.0}$ | **61.4**$_{\pm 1.0}$ | **95.3**$_{\pm 0.1}$ | 78.1$_{\pm 0.6}$ | 83.0$_{\pm 0.1}$ | **83.2** |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 91.2$_{\pm 0.4}$ | 98.4$_{\pm 0.4}$ | 95.5$_{\pm 0.7}$ | 85.9$_{\pm 0.5}$ | 73.7$_{\pm 2.5}$ | 88.9 |
| Energy | 87.1$_{\pm 9.2}$ | 97.3$_{\pm 1.3}$ | 93.1$_{\pm 4.8}$ | 85.3$_{\pm 0.7}$ | 71.6$_{\pm 2.0}$ | 86.9 |
| MLS | 88.6$_{\pm 7.0}$ | 98.0$_{\pm 0.8}$ | 94.1$_{\pm 3.7}$ | 84.7$_{\pm 0.7}$ | 71.8$_{\pm 1.3}$ | 87.4 |
| KLM | 92.9$_{\pm 0.5}$ | 98.4$_{\pm 0.5}$ | 95.5$_{\pm 0.8}$ | 84.1$_{\pm 0.6}$ | 80.6$_{\pm 2.7}$ | 90.3 |
| GNorm | 89.2$_{\pm 0.5}$ | 98.5$_{\pm 0.4}$ | 95.3$_{\pm 0.5}$ | 85.0$_{\pm 0.5}$ | **52.1**$_{\pm 0.9}$ | 84.0 |
| ReAct | 86.2$_{\pm 9.8}$ | 96.3$_{\pm 2.3}$ | 94.5$_{\pm 4.0}$ | 85.5$_{\pm 0.7}$ | 59.6$_{\pm 1.7}$ | 84.4 |
| VIM | 64.6$_{\pm 8.5}$ | 95.4$_{\pm 0.9}$ | 72.2$_{\pm 4.7}$ | 89.7$_{\pm 0.5}$ | 87.6$_{\pm 0.1}$ | 81.9 |
| MDS | 61.9$_{\pm 0.1}$ | 95.1$_{\pm 0.4}$ | 59.6$_{\pm 0.3}$ | 93.8$_{\pm 0.3}$ | 93.8$_{\pm 0.0}$ | 80.8 |
| RMDS | 94.2$_{\pm 0.1}$ | 99.1$_{\pm 0.1}$ | 96.4$_{\pm 0.1}$ | 83.5$_{\pm 0.9}$ | 88.0$_{\pm 0.1}$ | 92.2 |
| KNN | 9.3$_{\pm 0.1}$ | 94.9$_{\pm 0.8}$ | 50.2$_{\pm 0.3}$ | 83.9$_{\pm 0.9}$ | 73.6$_{\pm 0.1}$ | 62.4 |
| SHE | 32.7$_{\pm 0.3}$ | 96.6$_{\pm 0.2}$ | 60.5$_{\pm 0.7}$ | 85.1$_{\pm 0.4}$ | 53.4$_{\pm 0.8}$ | 65.7 |
| **SPROD** | **7.5**$_{\pm 0.1}$ | **94.5**$_{\pm 0.7}$ | **31.6**$_{\pm 0.5}$ | **78.4**$_{\pm 0.9}$ | 60.8$_{\pm 0.3}$ | **54.6** |

Table 17: AUROC and FPR@95 performance of all methods using ResNet-50 as the feature backbone.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 62.3$_{\pm 0.6}$ | 46.0$_{\pm 1.4}$ | 38.5$_{\pm 0.3}$ | 79.7$_{\pm 0.4}$ | 83.1$_{\pm 0.3}$ | 61.9 |
| Energy | 62.0$_{\pm 2.6}$ | 45.4$_{\pm 3.4}$ | 38.4$_{\pm 2.1}$ | 79.9$_{\pm 0.6}$ | 80.6$_{\pm 0.4}$ | 61.3 |
| MLS | 62.2$_{\pm 2.3}$ | 45.3$_{\pm 3.2}$ | 38.4$_{\pm 1.4}$ | 80.2$_{\pm 0.6}$ | 81.9$_{\pm 0.3}$ | 61.6 |
| KLM | 51.2$_{\pm 0.7}$ | 41.7$_{\pm 2.5}$ | 57.0$_{\pm 0.2}$ | 74.2$_{\pm 0.6}$ | 79.6$_{\pm 0.8}$ | 60.7 |
| GNorm | 79.5$_{\pm 0.4}$ | 38.0$_{\pm 1.3}$ | 46.6$_{\pm 0.4}$ | 74.2$_{\pm 0.5}$ | 85.2$_{\pm 0.2}$ | 64.7 |
| ReAct | 72.9$_{\pm 3.6}$ | 45.6$_{\pm 5.3}$ | 41.3$_{\pm 3.1}$ | 80.1$_{\pm 0.6}$ | 83.6$_{\pm 0.7}$ | 64.7 |
| VIM | 79.6$_{\pm 2.5}$ | 50.4$_{\pm 3.1}$ | 60.7$_{\pm 1.7}$ | 78.6$_{\pm 0.6}$ | 77.4$_{\pm 0.9}$ | 69.3 |
| MDS | 90.2$_{\pm 0.1}$ | 57.8$_{\pm 0.5}$ | 91.8$_{\pm 0.1}$ | 62.9$_{\pm 0.8}$ | 58.4$_{\pm 0.1}$ | 72.2 |
| RMDS | 59.4$_{\pm 0.1}$ | 33.6$_{\pm 1.4}$ | 47.4$_{\pm 0.2}$ | 81.9$_{\pm 0.4}$ | 68.8$_{\pm 0.1}$ | 58.2 |
| KNN | 98.6$_{\pm 0.0}$ | 54.5$_{\pm 0.5}$ | 91.1$_{\pm 0.1}$ | 79.7$_{\pm 0.6}$ | 77.4$_{\pm 0.0}$ | 80.3 |
| SHE | 88.3$_{\pm 0.2}$ | 42.7$_{\pm 0.6}$ | 73.2$_{\pm 0.1}$ | 54.8$_{\pm 0.7}$ | 83.0$_{\pm 0.1}$ | 68.4 |
| **SPROD** | **98.8**$_{\pm 0.0}$ | **61.6**$_{\pm 0.9}$ | **97.4**$_{\pm 0.0}$ | **82.4**$_{\pm 0.5}$ | **85.3**$_{\pm 0.0}$ | 85.1 |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 87.9$_{\pm 0.8}$ | 98.7$_{\pm 0.5}$ | 97.3$_{\pm 0.3}$ | 83.8$_{\pm 0.7}$ | 74.1$_{\pm 1.2}$ | 88.4 |
| Energy | 89.2$_{\pm 3.2}$ | 98.6$_{\pm 0.7}$ | 95.5$_{\pm 3.1}$ | 84.8$_{\pm 0.8}$ | 76.3$_{\pm 0.9}$ | 88.9 |
| MLS | 88.1$_{\pm 2.0}$ | 98.8$_{\pm 0.6}$ | 96.7$_{\pm 2.0}$ | 84.4$_{\pm 0.8}$ | 74.6$_{\pm 0.9}$ | 88.5 |
| KLM | 89.1$_{\pm 0.7}$ | 98.7$_{\pm 0.5}$ | 97.1$_{\pm 0.3}$ | 80.5$_{\pm 0.8}$ | 76.1$_{\pm 1.7}$ | 88.3 |
| GNorm | 84.2$_{\pm 0.7}$ | 98.8$_{\pm 0.4}$ | 97.1$_{\pm 0.1}$ | 84.2$_{\pm 0.6}$ | 54.7$_{\pm 0.6}$ | 83.8 |
| ReAct | 86.9$_{\pm 7.0}$ | 96.3$_{\pm 2.4}$ | 95.5$_{\pm 3.2}$ | 83.9$_{\pm 0.8}$ | 57.5$_{\pm 1.6}$ | 84.0 |
| VIM | 61.4$_{\pm 3.5}$ | 96.2$_{\pm 0.4}$ | 69.0$_{\pm 1.5}$ | 86.6$_{\pm 0.7}$ | 79.5$_{\pm 0.5}$ | 78.5 |
| MDS | 49.2$_{\pm 0.2}$ | 96.0$_{\pm 0.5}$ | 39.0$_{\pm 0.3}$ | 93.0$_{\pm 0.3}$ | 90.5$_{\pm 0.1}$ | 73.5 |
| RMDS | 91.7$_{\pm 0.2}$ | 99.6$_{\pm 0.1}$ | 95.3$_{\pm 0.1}$ | 83.4$_{\pm 0.9}$ | 88.1$_{\pm 0.1}$ | 91.6 |
| KNN | 4.8$_{\pm 0.1}$ | 94.4$_{\pm 1.0}$ | 42.5$_{\pm 0.2}$ | 79.9$_{\pm 1.1}$ | 70.4$_{\pm 0.2}$ | 58.4 |
| SHE | 33.2$_{\pm 0.5}$ | 96.4$_{\pm 0.5}$ | 76.5$_{\pm 0.2}$ | 93.9$_{\pm 0.3}$ | **52.6**$_{\pm 0.8}$ | 70.5 |
| **SPROD** | **4.7**$_{\pm 0.1}$ | **93.7**$_{\pm 0.9}$ | **19.0**$_{\pm 0.4}$ | **69.5**$_{\pm 1.2}$ | 58.0$_{\pm 0.1}$ | **49.0** |

Table 18: AUROC and FPR@95 performance of all methods using ResNet-101 as the feature backbone.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 65.5$_{\pm 0.6}$ | 44.9$_{\pm 3.1}$ | 41.3$_{\pm 1.2}$ | 81.2$_{\pm 0.5}$ | 83.8$_{\pm 0.2}$ | 63.3 |
| Energy | 63.9$_{\pm 2.9}$ | 45.5$_{\pm 3.9}$ | 41.1$_{\pm 2.5}$ | 81.8$_{\pm 0.7}$ | 81.3$_{\pm 0.4}$ | 62.7 |
| MLS | 64.3$_{\pm 2.3}$ | 45.3$_{\pm 3.7}$ | 41.2$_{\pm 2.1}$ | 82.0$_{\pm 0.7}$ | 82.7$_{\pm 0.2}$ | 63.1 |
| KLM | 49.7$_{\pm 0.7}$ | 45.0$_{\pm 2.5}$ | 56.4$_{\pm 1.2}$ | 76.1$_{\pm 0.6}$ | 80.1$_{\pm 0.3}$ | 61.5 |
| GNorm | 76.5$_{\pm 0.5}$ | 45.4$_{\pm 1.8}$ | 50.0$_{\pm 0.9}$ | 75.1$_{\pm 0.6}$ | 85.0$_{\pm 0.2}$ | 66.4 |
| ReAct | 72.0$_{\pm 3.7}$ | 39.0$_{\pm 7.9}$ | 43.5$_{\pm 3.8}$ | 81.3$_{\pm 0.9}$ | 84.9$_{\pm 0.7}$ | 64.1 |
| VIM | 88.5$_{\pm 1.7}$ | 48.3$_{\pm 3.7}$ | 63.8$_{\pm 1.4}$ | 79.4$_{\pm 0.7}$ | 80.5$_{\pm 0.6}$ | 72.1 |
| MDS | 96.1$_{\pm 0.0}$ | 53.4$_{\pm 0.4}$ | 93.2$_{\pm 0.0}$ | 63.7$_{\pm 0.9}$ | 64.7$_{\pm 0.1}$ | 74.2 |
| RMDS | 60.4$_{\pm 0.1}$ | 31.1$_{\pm 0.4}$ | 43.2$_{\pm 0.1}$ | **83.8**$_{\pm 0.5}$ | 73.2$_{\pm 0.1}$ | 58.3 |
| KNN | 98.9$_{\pm 0.0}$ | 49.6$_{\pm 5.3}$ | 89.5$_{\pm 0.1}$ | 81.5$_{\pm 0.6}$ | 80.2$_{\pm 0.1}$ | 79.9 |
| SHE | 83.5$_{\pm 0.4}$ | **64.6**$_{\pm 1.0}$ | 80.3$_{\pm 0.2}$ | 57.9$_{\pm 1.1}$ | 81.0$_{\pm 0.2}$ | 73.5 |
| **SPROD** | **99.0**$_{\pm 0.0}$ | 62.8$_{\pm 1.2}$ | **97.9**$_{\pm 0.0}$ | 83.1$_{\pm 0.5}$ | **86.8**$_{\pm 0.1}$ | **85.9** |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 87.7$_{\pm 0.7}$ | 98.9$_{\pm 0.3}$ | 95.9$_{\pm 0.5}$ | 81.0$_{\pm 0.9}$ | 69.0$_{\pm 1.1}$ | 86.5 |
| Energy | 90.4$_{\pm 5.6}$ | 98.2$_{\pm 0.9}$ | 95.2$_{\pm 2.2}$ | 79.2$_{\pm 1.2}$ | 74.4$_{\pm 1.0}$ | 87.5 |
| MLS | 89.4$_{\pm 3.2}$ | 98.8$_{\pm 0.4}$ | 95.7$_{\pm 1.6}$ | 79.2$_{\pm 1.1}$ | 70.9$_{\pm 0.7}$ | 86.8 |
| KLM | 89.0$_{\pm 0.7}$ | 98.9$_{\pm 0.3}$ | 95.8$_{\pm 0.5}$ | 78.6$_{\pm 0.9}$ | 74.5$_{\pm 1.1}$ | 87.4 |
| GNorm | 85.7$_{\pm 0.9}$ | 98.8$_{\pm 0.4}$ | 95.7$_{\pm 0.6}$ | 80.2$_{\pm 0.8}$ | 53.9$_{\pm 1.3}$ | 82.9 |
| ReAct | 86.9$_{\pm 8.7}$ | 97.2$_{\pm 1.7}$ | 94.6$_{\pm 2.9}$ | 79.9$_{\pm 1.1}$ | 59.0$_{\pm 2.2}$ | 83.5 |
| VIM | 42.2$_{\pm 4.2}$ | 97.2$_{\pm 0.6}$ | 68.3$_{\pm 2.6}$ | 87.0$_{\pm 0.8}$ | 76.9$_{\pm 0.9}$ | 74.3 |
| MDS | 23.6$_{\pm 0.2}$ | 95.8$_{\pm 0.6}$ | 40.4$_{\pm 0.3}$ | 92.9$_{\pm 0.3}$ | 85.8$_{\pm 0.1}$ | 67.7 |
| RMDS | 89.3$_{\pm 0.2}$ | 99.5$_{\pm 0.1}$ | 96.0$_{\pm 0.1}$ | 79.4$_{\pm 1.2}$ | 82.7$_{\pm 0.2}$ | 89.4 |
| KNN | **3.1**$_{\pm 0.1}$ | 94.9$_{\pm 1.3}$ | 43.2$_{\pm 0.2}$ | 76.8$_{\pm 1.3}$ | 68.3$_{\pm 0.1}$ | 57.3 |
| SHE | 41.5$_{\pm 0.6}$ | **89.0**$_{\pm 1.1}$ | 59.4$_{\pm 0.7}$ | 90.1$_{\pm 0.7}$ | 55.0$_{\pm 0.4}$ | 67.0 |
| **SPROD** | 3.3$_{\pm 0.1}$ | 94.8$_{\pm 1.1}$ | **11.4**$_{\pm 0.2}$ | **70.6**$_{\pm 1.2}$ | **52.9**$_{\pm 0.2}$ | **46.6** |

Table 19: AUROC and FPR@95 performance of all methods using DINOv2-S as the feature backbone.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 80.1$_{\pm 2.0}$ | 34.0$_{\pm 3.3}$ | 64.3$_{\pm 0.5}$ | 87.6$_{\pm 0.4}$ | 84.9$_{\pm 0.4}$ | 70.2 |
| Energy | 77.4$_{\pm 3.0}$ | 36.5$_{\pm 9.2}$ | 63.2$_{\pm 9.9}$ | 88.8$_{\pm 0.4}$ | 86.6$_{\pm 0.8}$ | 70.5 |
| MLS | 78.0$_{\pm 2.6}$ | 35.9$_{\pm 8.9}$ | 63.5$_{\pm 9.1}$ | 88.8$_{\pm 0.4}$ | 86.5$_{\pm 0.6}$ | 70.5 |
| KLM | 64.7$_{\pm 1.4}$ | 54.1$_{\pm 3.6}$ | 65.2$_{\pm 6.3}$ | 82.8$_{\pm 0.7}$ | 84.4$_{\pm 0.3}$ | 70.2 |
| GNorm | 82.7$_{\pm 1.5}$ | 32.8$_{\pm 2.4}$ | 63.5$_{\pm 0.5}$ | 83.4$_{\pm 0.7}$ | **86.9**$_{\pm 0.6}$ | 69.9 |
| ReAct | 76.5$_{\pm 2.8}$ | 35.2$_{\pm 8.2}$ | 62.9$_{\pm 9.6}$ | 88.7$_{\pm 0.4}$ | 86.7$_{\pm 0.7}$ | 70.0 |
| VIM | 97.8$_{\pm 0.4}$ | 39.0$_{\pm 8.2}$ | 83.0$_{\pm 7.2}$ | 86.5$_{\pm 0.7}$ | 84.9$_{\pm 0.3}$ | 78.2 |
| MDS | 99.1$_{\pm 0.0}$ | 55.0$_{\pm 0.7}$ | 94.1$_{\pm 0.0}$ | 83.8$_{\pm 0.8}$ | 83.7$_{\pm 0.1}$ | 83.1 |
| RMDS | 65.2$_{\pm 0.1}$ | 26.4$_{\pm 0.9}$ | 50.3$_{\pm 0.2}$ | 89.8$_{\pm 0.6}$ | 85.0$_{\pm 0.1}$ | 63.3 |
| KNN | 99.3$_{\pm 0.0}$ | 43.4$_{\pm 0.5}$ | 83.9$_{\pm 0.4}$ | **91.2**$_{\pm 0.3}$ | 80.5$_{\pm 0.8}$ | 79.7 |
| SHE | 99.1$_{\pm 0.0}$ | 54.6$_{\pm 0.9}$ | 92.5$_{\pm 0.1}$ | 84.3$_{\pm 1.1}$ | 84.7$_{\pm 0.2}$ | 83.0 |
| **SPROD** | **99.6**$_{\pm 0.0}$ | **63.4**$_{\pm 1.0}$ | **96.1**$_{\pm 0.0}$ | **91.2**$_{\pm 0.3}$ | 85.5$_{\pm 0.1}$ | **87.2** |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 80.1$_{\pm 2.6}$ | 99.0$_{\pm 0.2}$ | 91.8$_{\pm 1.6}$ | 71.6$_{\pm 1.2}$ | 63.0$_{\pm 2.8}$ | 81.1 |
| Energy | 79.0$_{\pm 7.5}$ | 97.9$_{\pm 1.5}$ | 87.5$_{\pm 7.9}$ | 69.3$_{\pm 1.7}$ | 56.0$_{\pm 3.7}$ | 77.9 |
| MLS | 78.2$_{\pm 7.2}$ | 98.4$_{\pm 1.0}$ | 88.0$_{\pm 7.6}$ | 69.3$_{\pm 1.6}$ | 57.8$_{\pm 3.1}$ | 78.3 |
| KLM | 82.7$_{\pm 2.7}$ | 99.0$_{\pm 0.2}$ | 93.2$_{\pm 3.2}$ | 63.7$_{\pm 1.2}$ | 64.4$_{\pm 2.2}$ | 80.6 |
| GNorm | 79.6$_{\pm 2.7}$ | 99.0$_{\pm 0.2}$ | 91.9$_{\pm 1.6}$ | 73.1$_{\pm 1.4}$ | 57.3$_{\pm 2.3}$ | 80.2 |
| ReAct | 79.8$_{\pm 7.1}$ | 97.8$_{\pm 1.5}$ | 87.7$_{\pm 7.5}$ | 68.9$_{\pm 1.7}$ | 57.2$_{\pm 3.4}$ | 78.3 |
| VIM | 6.0$_{\pm 1.7}$ | 97.4$_{\pm 1.1}$ | 56.6$_{\pm 13.1}$ | 67.1$_{\pm 2.2}$ | 56.6$_{\pm 1.1}$ | 56.7 |
| MDS | 2.5$_{\pm 0.1}$ | 96.4$_{\pm 0.4}$ | 30.4$_{\pm 0.2}$ | 69.2$_{\pm 2.3}$ | 58.6$_{\pm 0.2}$ | 51.4 |
| RMDS | 85.9$_{\pm 0.1}$ | 99.8$_{\pm 0.1}$ | 94.8$_{\pm 0.1}$ | 59.7$_{\pm 2.2}$ | 64.6$_{\pm 0.2}$ | 81.0 |
| KNN | 1.6$_{\pm 0.1}$ | 96.9$_{\pm 0.4}$ | 44.3$_{\pm 1.4}$ | 48.8$_{\pm 2.1}$ | 67.6$_{\pm 2.5}$ | 51.8 |
| SHE | 2.9$_{\pm 0.1}$ | 98.2$_{\pm 0.2}$ | 35.4$_{\pm 0.3}$ | 72.7$_{\pm 1.2}$ | 61.9$_{\pm 0.8}$ | 54.2 |
| **SPROD** | **1.3**$_{\pm 0.0}$ | **94.7**$_{\pm 0.7}$ | **24.4**$_{\pm 0.2}$ | **40.0**$_{\pm 1.3}$ | **54.4**$_{\pm 0.2}$ | **43.0** |

Table 20: AUROC and FPR@95 performance of all methods using ViT-S as the feature backbone.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 71.5$_{\pm 1.9}$ | 39.2$_{\pm 1.2}$ | 47.8$_{\pm 0.4}$ | 83.2$_{\pm 0.2}$ | 85.2$_{\pm 0.3}$ | 65.4 |
| Energy | 69.0$_{\pm 9.8}$ | 40.3$_{\pm 10.5}$ | 48.9$_{\pm 5.0}$ | 85.4$_{\pm 0.4}$ | 85.6$_{\pm 0.7}$ | 65.8 |
| MLS | 69.3$_{\pm 9.1}$ | 39.8$_{\pm 9.9}$ | 48.7$_{\pm 4.6}$ | 85.3$_{\pm 0.4}$ | 85.7$_{\pm 0.5}$ | 65.8 |
| KLM | 62.6$_{\pm 2.5}$ | 50.6$_{\pm 4.7}$ | 47.2$_{\pm 5.6}$ | 77.9$_{\pm 0.5}$ | 84.4$_{\pm 0.6}$ | 64.5 |
| GNorm | 76.0$_{\pm 1.4}$ | 38.4$_{\pm 1.0}$ | 47.9$_{\pm 0.6}$ | 83.0$_{\pm 0.4}$ | **86.7**$_{\pm 0.3}$ | 66.4 |
| ReAct | 70.3$_{\pm 7.8}$ | 49.1$_{\pm 9.6}$ | 49.0$_{\pm 5.2}$ | 84.9$_{\pm 0.4}$ | 83.4$_{\pm 1.0}$ | 67.3 |
| VIM | 84.5$_{\pm 7.9}$ | 44.0$_{\pm 9.7}$ | 67.8$_{\pm 6.0}$ | 85.9$_{\pm 0.4}$ | 85.5$_{\pm 0.5}$ | 73.5 |
| MDS | 95.1$_{\pm 0.0}$ | 59.6$_{\pm 1.0}$ | 94.3$_{\pm 0.0}$ | 81.6$_{\pm 0.5}$ | 83.9$_{\pm 0.0}$ | 82.9 |
| RMDS | 60.5$_{\pm 0.1}$ | 29.2$_{\pm 1.0}$ | 42.1$_{\pm 0.1}$ | **87.2**$_{\pm 0.3}$ | 86.4$_{\pm 0.1}$ | 61.1 |
| KNN | 94.0$_{\pm 0.0}$ | 53.2$_{\pm 4.1}$ | 91.2$_{\pm 0.0}$ | 83.3$_{\pm 0.5}$ | 79.6$_{\pm 0.1}$ | 80.3 |
| SHE | 94.5$_{\pm 0.2}$ | 51.5$_{\pm 0.6}$ | 91.2$_{\pm 0.2}$ | 78.3$_{\pm 2.0}$ | 83.4$_{\pm 0.2}$ | 79.8 |
| **SPROD** | **97.0**$_{\pm 0.0}$ | **60.9**$_{\pm 1.1}$ | **96.3**$_{\pm 0.0}$ | 85.0$_{\pm 0.3}$ | 86.1$_{\pm 0.1}$ | **85.1** |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 88.7$_{\pm 1.2}$ | 99.0$_{\pm 0.2}$ | 96.0$_{\pm 0.6}$ | 81.3$_{\pm 0.6}$ | 67.1$_{\pm 1.6}$ | 86.4 |
| Energy | 87.9$_{\pm 8.4}$ | 98.3$_{\pm 1.5}$ | 93.6$_{\pm 3.7}$ | 78.2$_{\pm 1.2}$ | 66.5$_{\pm 2.1}$ | 84.9 |
| MLS | 88.3$_{\pm 7.3}$ | 98.6$_{\pm 1.0}$ | 93.9$_{\pm 3.0}$ | 78.6$_{\pm 1.1}$ | 65.0$_{\pm 1.4}$ | 84.9 |
| KLM | 89.7$_{\pm 2.1}$ | 99.1$_{\pm 0.2}$ | 96.2$_{\pm 2.5}$ | 78.3$_{\pm 0.7}$ | 65.4$_{\pm 2.1}$ | 85.7 |
| GNorm | 88.4$_{\pm 1.2}$ | 99.0$_{\pm 0.2}$ | 95.9$_{\pm 0.5}$ | 80.6$_{\pm 0.8}$ | 62.5$_{\pm 0.7}$ | 85.3 |
| ReAct | 86.4$_{\pm 11.2}$ | 97.0$_{\pm 2.9}$ | 94.1$_{\pm 3.0}$ | 79.6$_{\pm 1.1}$ | 76.7$_{\pm 2.3}$ | 86.8 |
| VIM | 63.6$_{\pm 15.1}$ | 97.5$_{\pm 1.4}$ | 78.1$_{\pm 6.7}$ | 72.1$_{\pm 1.2}$ | 63.8$_{\pm 1.3}$ | 75.0 |
| MDS | 30.9$_{\pm 0.1}$ | 95.6$_{\pm 0.3}$ | 42.3$_{\pm 0.2}$ | 75.6$_{\pm 1.2}$ | 65.1$_{\pm 0.2}$ | 61.9 |
| RMDS | 91.2$_{\pm 0.1}$ | 99.8$_{\pm 0.1}$ | 96.8$_{\pm 0.1}$ | 71.0$_{\pm 1.7}$ | 60.1$_{\pm 0.1}$ | 83.8 |
| KNN | 39.7$_{\pm 0.1}$ | 95.8$_{\pm 0.5}$ | 50.1$_{\pm 0.1}$ | 75.4$_{\pm 1.3}$ | 72.7$_{\pm 0.2}$ | 66.7 |
| SHE | 28.6$_{\pm 1.8}$ | 98.2$_{\pm 0.2}$ | 42.2$_{\pm 1.6}$ | 80.2$_{\pm 1.6}$ | 66.3$_{\pm 1.4}$ | 63.1 |
| **SPROD** | **9.1**$_{\pm 0.2}$ | **95.0**$_{\pm 0.9}$ | **25.0**$_{\pm 0.1}$ | **66.6**$_{\pm 1.1}$ | **60.0**$_{\pm 0.2}$ | **51.1** |

Table 21: AUROC and FPR@95 performance of all methods using Swin-B as the feature backbone.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 82.8$_{\pm 0.9}$ | 30.9$_{\pm 2.0}$ | 57.2$_{\pm 0.4}$ | 92.7$_{\pm 0.2}$ | 90.6$_{\pm 0.4}$ | 70.8 |
| Energy | 82.4$_{\pm 1.5}$ | 31.0$_{\pm 2.9}$ | 57.5$_{\pm 2.7}$ | **93.6**$_{\pm 0.3}$ | 87.6$_{\pm 0.6}$ | 70.4 |
| MLS | 82.6$_{\pm 1.3}$ | 30.9$_{\pm 2.8}$ | 57.5$_{\pm 2.2}$ | 92.9$_{\pm 0.3}$ | 89.6$_{\pm 0.4}$ | 70.7 |
| KLM | 57.3$_{\pm 1.7}$ | 52.8$_{\pm 3.5}$ | 50.5$_{\pm 0.5}$ | 92.5$_{\pm 0.3}$ | 90.1$_{\pm 0.3}$ | 68.6 |
| GNorm | 63.4$_{\pm 1.1}$ | 27.6$_{\pm 1.2}$ | 52.4$_{\pm 0.4}$ | 84.0$_{\pm 0.9}$ | 56.0$_{\pm 0.8}$ | 56.7 |
| ReAct | 83.6$_{\pm 1.7}$ | 34.3$_{\pm 2.9}$ | 58.0$_{\pm 2.6}$ | **93.6**$_{\pm 0.3}$ | 90.8$_{\pm 0.4}$ | 72.1 |
| VIM | 90.6$_{\pm 0.3}$ | 39.8$_{\pm 3.2}$ | 76.9$_{\pm 1.4}$ | 88.1$_{\pm 0.3}$ | 92.6$_{\pm 0.1}$ | 77.6 |
| MDS | 85.9$_{\pm 0.1}$ | 60.8$_{\pm 1.0}$ | 86.7$_{\pm 0.2}$ | 74.7$_{\pm 0.8}$ | 92.1$_{\pm 0.0}$ | 80.0 |
| RMDS | 82.6$_{\pm 0.1}$ | 29.9$_{\pm 1.3}$ | 49.5$_{\pm 0.1}$ | 93.2$_{\pm 0.2}$ | 92.8$_{\pm 0.0}$ | 69.6 |
| KNN | 90.4$_{\pm 0.0}$ | 55.4$_{\pm 0.8}$ | 91.8$_{\pm 0.2}$ | 92.5$_{\pm 0.3}$ | 91.3$_{\pm 0.0}$ | 84.3 |
| SHE | 97.3$_{\pm 0.0}$ | 49.6$_{\pm 0.9}$ | 96.4$_{\pm 0.1}$ | 92.9$_{\pm 0.4}$ | 88.2$_{\pm 0.1}$ | 84.9 |
| **SPROD** | **99.0**$_{\pm 0.0}$ | **66.6**$_{\pm 1.2}$ | **98.7**$_{\pm 0.0}$ | 92.7$_{\pm 0.2}$ | **93.7**$_{\pm 0.0}$ | **90.1** |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 76.1$_{\pm 2.7}$ | 99.7$_{\pm 0.1}$ | 93.0$_{\pm 0.6}$ | 51.8$_{\pm 2.0}$ | 38.1$_{\pm 2.0}$ | 71.7 |
| Energy | 76.3$_{\pm 6.3}$ | 99.4$_{\pm 0.3}$ | 91.9$_{\pm 3.2}$ | 47.3$_{\pm 2.3}$ | 53.5$_{\pm 1.8}$ | 73.7 |
| MLS | 75.7$_{\pm 4.7}$ | 99.6$_{\pm 0.2}$ | 92.9$_{\pm 2.6}$ | 48.5$_{\pm 2.2}$ | 41.0$_{\pm 1.4}$ | 71.5 |
| KLM | 79.3$_{\pm 2.1}$ | 99.7$_{\pm 0.1}$ | 92.9$_{\pm 0.8}$ | **41.3**$_{\pm 1.7}$ | 49.8$_{\pm 5.3}$ | 72.6 |
| GNorm | 75.9$_{\pm 2.4}$ | 99.7$_{\pm 0.1}$ | 93.1$_{\pm 0.6}$ | 46.5$_{\pm 2.2}$ | 80.9$_{\pm 2.2}$ | 79.2 |
| ReAct | 75.9$_{\pm 6.4}$ | 99.0$_{\pm 0.4}$ | 91.8$_{\pm 2.6}$ | 48.0$_{\pm 2.2}$ | 48.3$_{\pm 1.4}$ | 72.6 |
| VIM | 58.2$_{\pm 2.0}$ | 97.7$_{\pm 0.6}$ | 65.0$_{\pm 1.9}$ | 82.7$_{\pm 1.2}$ | 37.8$_{\pm 0.6}$ | 68.3 |
| MDS | 67.5$_{\pm 0.2}$ | 94.3$_{\pm 1.0}$ | 57.3$_{\pm 0.5}$ | 92.9$_{\pm 0.6}$ | 41.5$_{\pm 0.1}$ | 70.7 |
| RMDS | 67.4$_{\pm 0.1}$ | 99.9$_{\pm 0.0}$ | 95.4$_{\pm 0.2}$ | 64.3$_{\pm 2.3}$ | 37.7$_{\pm 0.2}$ | 72.9 |
| KNN | 43.5$_{\pm 0.1}$ | 93.9$_{\pm 0.5}$ | 42.2$_{\pm 0.4}$ | 47.9$_{\pm 2.0}$ | 39.7$_{\pm 0.2}$ | 53.4 |
| SHE | 16.4$_{\pm 0.4}$ | 98.6$_{\pm 0.2}$ | 24.1$_{\pm 0.8}$ | 50.4$_{\pm 3.7}$ | 42.4$_{\pm 0.5}$ | 46.4 |
| **SPROD** | **2.8**$_{\pm 0.1}$ | **92.3**$_{\pm 0.9}$ | **8.4**$_{\pm 0.2}$ | 43.6$_{\pm 1.2}$ | **29.6**$_{\pm 0.2}$ | **35.3** |

Table 22: AUROC and FPR@95 performance of all methods using DeiT-B as the feature backbone.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 75.6$_{\pm 0.9}$ | 43.4$_{\pm 2.2}$ | 53.3$_{\pm 0.1}$ | 88.5$_{\pm 0.4}$ | 87.8$_{\pm 0.4}$ | 69.7 |
| Energy | 77.2$_{\pm 3.1}$ | 44.3$_{\pm 4.0}$ | 52.4$_{\pm 3.1}$ | 90.0$_{\pm 0.6}$ | 86.2$_{\pm 1.1}$ | 70.0 |
| MLS | 76.9$_{\pm 2.5}$ | 44.0$_{\pm 3.7}$ | 52.6$_{\pm 2.5}$ | 89.6$_{\pm 0.6}$ | 87.4$_{\pm 0.4}$ | 70.1 |
| KLM | 47.0$_{\pm 0.6}$ | 43.2$_{\pm 2.0}$ | 49.6$_{\pm 0.3}$ | 85.8$_{\pm 0.4}$ | 86.6$_{\pm 0.4}$ | 62.4 |
| GNorm | 67.3$_{\pm 0.8}$ | 43.2$_{\pm 1.6}$ | 53.3$_{\pm 0.2}$ | 76.5$_{\pm 1.2}$ | 70.3$_{\pm 2.5}$ | 62.1 |
| ReAct | 77.1$_{\pm 3.0}$ | 41.9$_{\pm 4.0}$ | 52.7$_{\pm 3.3}$ | **90.2**$_{\pm 0.6}$ | 87.3$_{\pm 1.0}$ | 69.8 |
| VIM | 82.7$_{\pm 1.7}$ | 46.2$_{\pm 3.3}$ | 65.6$_{\pm 2.4}$ | 82.5$_{\pm 0.4}$ | 87.4$_{\pm 0.2}$ | 72.9 |
| MDS | 74.9$_{\pm 0.1}$ | 50.8$_{\pm 0.6}$ | 78.2$_{\pm 0.1}$ | 65.6$_{\pm 0.9}$ | 87.2$_{\pm 0.1}$ | 71.3 |
| RMDS | 69.2$_{\pm 0.1}$ | 31.7$_{\pm 0.3}$ | 50.0$_{\pm 0.3}$ | 87.3$_{\pm 0.3}$ | 89.1$_{\pm 0.1}$ | 65.5 |
| KNN | 90.0$_{\pm 0.6}$ | 46.4$_{\pm 4.3}$ | 87.6$_{\pm 0.0}$ | 83.9$_{\pm 0.6}$ | 86.5$_{\pm 0.1}$ | 78.9 |
| SHE | 93.9$_{\pm 0.0}$ | 54.1$_{\pm 0.7}$ | 94.0$_{\pm 0.0}$ | 80.1$_{\pm 1.3}$ | 86.9$_{\pm 0.1}$ | 81.8 |
| **SPROD** | **94.1**$_{\pm 0.0}$ | **54.8**$_{\pm 0.6}$ | **95.8**$_{\pm 0.0}$ | 88.9$_{\pm 0.3}$ | **89.5**$_{\pm 0.1}$ | **84.6** |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 86.7$_{\pm 0.6}$ | 98.9$_{\pm 0.2}$ | 94.9$_{\pm 0.6}$ | 58.9$_{\pm 1.7}$ | 62.5$_{\pm 2.3}$ | 80.4 |
| Energy | 80.3$_{\pm 8.6}$ | 98.2$_{\pm 1.0}$ | 94.3$_{\pm 3.3}$ | **51.5**$_{\pm 2.3}$ | 68.6$_{\pm 4.8}$ | 78.6 |
| MLS | 82.0$_{\pm 5.9}$ | 98.5$_{\pm 0.6}$ | 94.7$_{\pm 2.6}$ | 52.7$_{\pm 2.2}$ | 63.2$_{\pm 2.9}$ | 78.2 |
| KLM | 87.3$_{\pm 0.8}$ | 98.8$_{\pm 0.2}$ | 94.9$_{\pm 0.4}$ | 61.2$_{\pm 1.3}$ | 60.2$_{\pm 4.9}$ | 80.5 |
| GNorm | 84.4$_{\pm 0.7}$ | 98.9$_{\pm 0.2}$ | 94.6$_{\pm 0.4}$ | 52.5$_{\pm 2.2}$ | 81.0$_{\pm 4.6}$ | 82.3 |
| ReAct | 80.6$_{\pm 7.9}$ | 98.4$_{\pm 0.9}$ | 93.5$_{\pm 3.7}$ | 52.7$_{\pm 2.2}$ | 66.2$_{\pm 4.8}$ | 78.3 |
| VIM | 79.3$_{\pm 2.4}$ | 96.7$_{\pm 0.7}$ | 83.0$_{\pm 4.0}$ | 89.1$_{\pm 0.8}$ | 64.8$_{\pm 0.9}$ | 82.6 |
| MDS | 81.3$_{\pm 0.1}$ | 95.6$_{\pm 0.6}$ | 76.0$_{\pm 0.2}$ | 94.3$_{\pm 0.5}$ | 65.8$_{\pm 0.2}$ | 82.6 |
| RMDS | 86.9$_{\pm 0.1}$ | 99.5$_{\pm 0.1}$ | 95.1$_{\pm 0.1}$ | 86.6$_{\pm 0.8}$ | 58.0$_{\pm 0.2}$ | 85.2 |
| KNN | 58.6$_{\pm 2.0}$ | **95.3**$_{\pm 0.7}$ | 57.6$_{\pm 0.2}$ | 76.0$_{\pm 1.7}$ | 71.9$_{\pm 0.1}$ | 71.9 |
| SHE | 44.0$_{\pm 0.5}$ | 98.3$_{\pm 0.3}$ | 38.1$_{\pm 0.8}$ | 92.1$_{\pm 1.2}$ | 62.1$_{\pm 1.3}$ | 66.9 |
| **SPROD** | **42.3**$_{\pm 0.2}$ | **95.3**$_{\pm 0.6}$ | **30.6**$_{\pm 0.2}$ | 53.4$_{\pm 1.1}$ | **54.1**$_{\pm 0.1}$ | **55.1** |

Table 23: AUROC and FPR@95 performance of all methods using ConvNeXt-B as the feature backbone.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 88.1$_{\pm 0.7}$ | 30.4$_{\pm 2.5}$ | 57.5$_{\pm 0.4}$ | 91.1$_{\pm 0.2}$ | 93.1$_{\pm 0.2}$ | 72.0 |
| Energy | 86.1$_{\pm 4.0}$ | 29.1$_{\pm 4.7}$ | 55.7$_{\pm 5.7}$ | 93.6$_{\pm 0.2}$ | 92.9$_{\pm 0.3}$ | 71.5 |
| MLS | 86.6$_{\pm 3.8}$ | 28.9$_{\pm 4.5}$ | 56.0$_{\pm 5.0}$ | 91.7$_{\pm 0.2}$ | 93.1$_{\pm 0.3}$ | 71.3 |
| KLM | 65.4$_{\pm 1.8}$ | 48.2$_{\pm 2.8}$ | 50.0$_{\pm 0.4}$ | 92.6$_{\pm 0.2}$ | 90.4$_{\pm 0.4}$ | 69.3 |
| GNorm | 71.6$_{\pm 1.2}$ | 47.0$_{\pm 1.8}$ | 46.1$_{\pm 0.5}$ | 85.0$_{\pm 0.7}$ | 89.4$_{\pm 0.4}$ | 67.8 |
| ReAct | 85.4$_{\pm 4.2}$ | 26.1$_{\pm 3.7}$ | 55.7$_{\pm 5.5}$ | 93.2$_{\pm 0.2}$ | 92.9$_{\pm 0.2}$ | 70.7 |
| VIM | 97.6$_{\pm 0.5}$ | 32.4$_{\pm 4.3}$ | 82.6$_{\pm 3.2}$ | **94.7**$_{\pm 0.2}$ | 92.7$_{\pm 0.1}$ | 80.0 |
| MDS | 97.7$_{\pm 0.0}$ | 52.3$_{\pm 1.0}$ | 95.7$_{\pm 0.1}$ | 88.4$_{\pm 0.4}$ | 89.7$_{\pm 0.0}$ | 84.8 |
| RMDS | 85.4$_{\pm 0.1}$ | 26.9$_{\pm 0.8}$ | 48.9$_{\pm 0.2}$ | 94.1$_{\pm 0.2}$ | 92.4$_{\pm 0.0}$ | 69.5 |
| KNN | 96.7$_{\pm 0.0}$ | 56.2$_{\pm 0.7}$ | 93.4$_{\pm 0.1}$ | 91.0$_{\pm 0.3}$ | 93.1$_{\pm 0.0}$ | 86.1 |
| SHE | 95.1$_{\pm 0.1}$ | 62.0$_{\pm 1.1}$ | 99.1$_{\pm 0.1}$ | 75.4$_{\pm 0.8}$ | 91.5$_{\pm 0.1}$ | 84.6 |
| **SPROD** | **99.1**$_{\pm 0.0}$ | **64.0**$_{\pm 0.9}$ | **99.7**$_{\pm 0.0}$ | 92.0$_{\pm 0.5}$ | **94.3**$_{\pm 0.0}$ | **89.8** |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 59.0$_{\pm 3.4}$ | 99.6$_{\pm 0.1}$ | 93.7$_{\pm 1.1}$ | 55.0$_{\pm 1.5}$ | 31.5$_{\pm 1.1}$ | 67.8 |
| Energy | 65.1$_{\pm 10.2}$ | 99.4$_{\pm 0.6}$ | 91.1$_{\pm 7.3}$ | 53.3$_{\pm 1.9}$ | 32.3$_{\pm 1.8}$ | 68.2 |
| MLS | 62.4$_{\pm 9.2}$ | 99.6$_{\pm 0.4}$ | 92.1$_{\pm 5.8}$ | 52.9$_{\pm 1.8}$ | 32.0$_{\pm 2.5}$ | 67.8 |
| KLM | 59.7$_{\pm 3.7}$ | 99.6$_{\pm 0.1}$ | 93.6$_{\pm 0.9}$ | 41.6$_{\pm 1.3}$ | 39.7$_{\pm 5.4}$ | 66.8 |
| GNorm | 61.9$_{\pm 2.8}$ | 99.6$_{\pm 0.1}$ | 94.4$_{\pm 0.8}$ | 60.5$_{\pm 1.7}$ | 37.0$_{\pm 2.4}$ | 70.7 |
| ReAct | 67.2$_{\pm 8.4}$ | 99.5$_{\pm 0.4}$ | 91.5$_{\pm 6.7}$ | 56.3$_{\pm 1.8}$ | 32.4$_{\pm 2.3}$ | 69.4 |
| VIM | 9.4$_{\pm 4.3}$ | 98.6$_{\pm 0.6}$ | 51.5$_{\pm 6.7}$ | **37.5**$_{\pm 2.0}$ | 39.8$_{\pm 2.8}$ | 47.4 |
| MDS | 10.0$_{\pm 0.1}$ | 96.5$_{\pm 1.0}$ | 28.3$_{\pm 0.4}$ | 63.8$_{\pm 2.0}$ | 53.2$_{\pm 0.2}$ | 50.4 |
| RMDS | 50.6$_{\pm 0.1}$ | 99.9$_{\pm 0.1}$ | 94.9$_{\pm 0.2}$ | 54.6$_{\pm 2.3}$ | 38.5$_{\pm 0.1}$ | 67.7 |
| KNN | 15.7$_{\pm 0.1}$ | 94.6$_{\pm 1.3}$ | 23.8$_{\pm 0.4}$ | 66.6$_{\pm 1.7}$ | 33.8$_{\pm 0.1}$ | 46.9 |
| SHE | 22.0$_{\pm 0.6}$ | 97.0$_{\pm 0.5}$ | 3.6$_{\pm 0.5}$ | 91.5$_{\pm 0.6}$ | 38.4$_{\pm 0.8}$ | 50.5 |
| **SPROD** | **2.6**$_{\pm 0.1}$ | **93.8**$_{\pm 1.7}$ | **0.5**$_{\pm 0.0}$ | 51.4$_{\pm 1.8}$ | **27.5**$_{\pm 0.1}$ | **35.2** |

Table 24: AUROC and FPR@95 performance of all methods using BiT-R50x1 as the feature backbone.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 69.5$_{\pm 1.7}$ | 41.8$_{\pm 5.4}$ | 46.8$_{\pm 1.0}$ | 84.3$_{\pm 0.2}$ | 82.5$_{\pm 0.3}$ | 65.0 |
| Energy | 66.8$_{\pm 2.3}$ | 43.1$_{\pm 9.4}$ | 44.2$_{\pm 7.0}$ | **87.6**$_{\pm 0.3}$ | 81.6$_{\pm 0.6}$ | 64.7 |
| MLS | 67.0$_{\pm 2.1}$ | 42.9$_{\pm 9.3}$ | 44.4$_{\pm 6.3}$ | 87.4$_{\pm 0.3}$ | 82.2$_{\pm 0.5}$ | 64.8 |
| KLM | 48.4$_{\pm 1.0}$ | 41.9$_{\pm 6.0}$ | 51.1$_{\pm 1.1}$ | 78.6$_{\pm 0.4}$ | 83.0$_{\pm 0.4}$ | 60.6 |
| GNorm | 49.7$_{\pm 1.5}$ | 35.7$_{\pm 2.8}$ | 60.6$_{\pm 1.6}$ | 71.3$_{\pm 0.6}$ | 53.9$_{\pm 1.5}$ | 54.2 |
| ReAct | 57.0$_{\pm 2.3}$ | 36.0$_{\pm 8.0}$ | 46.0$_{\pm 7.5}$ | 81.9$_{\pm 0.5}$ | 81.4$_{\pm 1.0}$ | 60.5 |
| VIM | 93.8$_{\pm 1.6}$ | 48.0$_{\pm 10.2}$ | 73.3$_{\pm 7.5}$ | 87.5$_{\pm 0.3}$ | **86.6**$_{\pm 0.2}$ | 77.8 |
| MDS | **99.1**$_{\pm 0.0}$ | 61.9$_{\pm 0.7}$ | 98.5$_{\pm 0.0}$ | 78.2$_{\pm 0.8}$ | 83.3$_{\pm 0.1}$ | 84.2 |
| RMDS | 70.0$_{\pm 0.1}$ | 37.2$_{\pm 2.1}$ | 55.4$_{\pm 0.3}$ | 85.4$_{\pm 0.3}$ | 84.5$_{\pm 0.1}$ | 66.5 |
| KNN | 98.5$_{\pm 0.0}$ | 49.5$_{\pm 4.0}$ | 95.4$_{\pm 0.0}$ | 85.2$_{\pm 0.5}$ | 79.3$_{\pm 0.1}$ | 81.6 |
| SHE | 80.0$_{\pm 0.8}$ | 47.3$_{\pm 0.5}$ | 97.1$_{\pm 0.1}$ | 79.6$_{\pm 0.8}$ | 63.6$_{\pm 0.3}$ | 73.5 |
| **SPROD** | 98.5$_{\pm 0.0}$ | **67.1**$_{\pm 1.0}$ | **98.6**$_{\pm 0.0}$ | 87.1$_{\pm 0.4}$ | 84.3$_{\pm 0.1}$ | **87.1** |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 84.7$_{\pm 1.4}$ | 98.5$_{\pm 0.6}$ | 95.2$_{\pm 0.6}$ | 76.0$_{\pm 0.6}$ | 66.7$_{\pm 1.9}$ | 84.2 |
| Energy | 89.2$_{\pm 2.8}$ | 97.8$_{\pm 1.5}$ | 94.9$_{\pm 5.4}$ | 64.2$_{\pm 1.2}$ | 68.1$_{\pm 2.7}$ | 82.8 |
| MLS | 88.2$_{\pm 2.8}$ | 98.1$_{\pm 1.3}$ | 95.1$_{\pm 4.8}$ | 65.4$_{\pm 1.1}$ | 66.6$_{\pm 2.4}$ | 82.7 |
| KLM | 85.2$_{\pm 1.2}$ | 98.4$_{\pm 0.6}$ | 95.1$_{\pm 0.5}$ | 74.8$_{\pm 0.7}$ | 63.8$_{\pm 2.8}$ | 83.5 |
| GNorm | 85.1$_{\pm 1.2}$ | 98.6$_{\pm 0.6}$ | 95.0$_{\pm 0.7}$ | 70.1$_{\pm 0.9}$ | 89.6$_{\pm 1.6}$ | 87.7 |
| ReAct | 94.3$_{\pm 1.4}$ | 97.4$_{\pm 1.9}$ | 95.2$_{\pm 4.7}$ | 74.2$_{\pm 1.1}$ | 69.4$_{\pm 3.4}$ | 86.1 |
| VIM | 27.9$_{\pm 5.2}$ | 97.2$_{\pm 1.3}$ | 61.0$_{\pm 12.0}$ | 72.0$_{\pm 1.4}$ | **51.0**$_{\pm 0.5}$ | 61.8 |
| MDS | **2.5**$_{\pm 0.1}$ | 95.7$_{\pm 0.5}$ | **5.9**$_{\pm 0.1}$ | 83.4$_{\pm 1.0}$ | 54.6$_{\pm 0.2}$ | 48.4 |
| RMDS | 83.3$_{\pm 0.1}$ | 98.5$_{\pm 0.4}$ | 90.4$_{\pm 0.1}$ | 82.1$_{\pm 1.2}$ | 56.8$_{\pm 0.2}$ | 82.2 |
| KNN | 3.7$_{\pm 0.1}$ | **95.5**$_{\pm 0.4}$ | 22.4$_{\pm 0.2}$ | 64.3$_{\pm 1.7}$ | 68.9$_{\pm 0.1}$ | 51.0 |
| SHE | 83.5$_{\pm 0.9}$ | 98.0$_{\pm 0.4}$ | 15.5$_{\pm 0.8}$ | 84.6$_{\pm 1.4}$ | 88.3$_{\pm 0.5}$ | 74.0 |
| **SPROD** | 4.9$_{\pm 0.1}$ | 95.6$_{\pm 0.5}$ | **5.9**$_{\pm 0.1}$ | **51.2**$_{\pm 1.1}$ | 56.8$_{\pm 0.3}$ | **42.9** |

Table 25: AUPR-IN and AUPR-OUT performance of all methods using ResNet-18 as the feature backbone.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 62.2$_{\pm 0.9}$ | 23.1$_{\pm 2.4}$ | 28.9$_{\pm 0.2}$ | 80.8$_{\pm 0.5}$ | 47.2$_{\pm 0.6}$ | 48.4 |
| Energy | 62.7$_{\pm 4.3}$ | 23.6$_{\pm 3.7}$ | 30.2$_{\pm 1.8}$ | 79.8$_{\pm 0.6}$ | 43.3$_{\pm 0.5}$ | 47.9 |
| MLS | 62.6$_{\pm 4.0}$ | 23.5$_{\pm 3.6}$ | 30.0$_{\pm 1.5}$ | 80.2$_{\pm 0.6}$ | 45.9$_{\pm 0.5}$ | 48.4 |
| KLM | 34.4$_{\pm 0.3}$ | 16.8$_{\pm 0.4}$ | 40.9$_{\pm 0.5}$ | 70.9$_{\pm 0.6}$ | 41.8$_{\pm 1.0}$ | 41.0 |
| GNorm | 76.0$_{\pm 0.4}$ | 16.7$_{\pm 0.8}$ | 35.1$_{\pm 0.2}$ | 73.7$_{\pm 0.4}$ | 50.0$_{\pm 0.5}$ | 50.3 |
| ReAct | 71.6$_{\pm 3.9}$ | 18.1$_{\pm 1.6}$ | 30.7$_{\pm 2.2}$ | 80.4$_{\pm 0.6}$ | **48.7**$_{\pm 0.8}$ | 49.9 |
| VIM | 71.5$_{\pm 3.5}$ | 28.3$_{\pm 4.1}$ | 42.1$_{\pm 3.0}$ | 80.6$_{\pm 0.5}$ | 35.5$_{\pm 0.5}$ | 51.6 |
| MDS | 61.6$_{\pm 0.1}$ | 29.9$_{\pm 0.5}$ | 86.8$_{\pm 0.2}$ | 67.0$_{\pm 0.9}$ | 15.9$_{\pm 0.0}$ | 52.2 |
| RMDS | 36.6$_{\pm 0.1}$ | 15.3$_{\pm 0.2}$ | 33.1$_{\pm 0.2}$ | **82.5**$_{\pm 0.4}$ | 43.7$_{\pm 0.1}$ | 42.2 |
| KNN | 88.1$_{\pm 0.2}$ | 27.4$_{\pm 0.5}$ | 88.7$_{\pm 1.4}$ | 80.3$_{\pm 0.5}$ | 33.4$_{\pm 0.1}$ | 63.6 |
| SHE | 81.6$_{\pm 0.2}$ | 17.5$_{\pm 0.1}$ | 57.1$_{\pm 0.2}$ | 60.1$_{\pm 0.7}$ | 47.9$_{\pm 0.3}$ | 52.8 |
| **SPROD** | **97.1**$_{\pm 0.1}$ | **40.4**$_{\pm 1.1}$ | **96.6**$_{\pm 0.1}$ | 81.0$_{\pm 0.5}$ | 48.4$_{\pm 0.1}$ | **72.7** |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 72.7$_{\pm 0.4}$ | 77.3$_{\pm 1.3}$ | 59.7$_{\pm 0.5}$ | 59.1$_{\pm 0.9}$ | 92.4$_{\pm 0.2}$ | 72.2 |
| Energy | 74.5$_{\pm 5.4}$ | 78.3$_{\pm 2.2}$ | 63.0$_{\pm 5.2}$ | 57.3$_{\pm 1.0}$ | 92.4$_{\pm 0.1}$ | 73.1 |
| MLS | 74.3$_{\pm 4.2}$ | 77.7$_{\pm 1.8}$ | 62.2$_{\pm 4.0}$ | 58.0$_{\pm 1.0}$ | 92.4$_{\pm 0.1}$ | 72.9 |
| KLM | 60.6$_{\pm 0.5}$ | 75.0$_{\pm 0.8}$ | 68.6$_{\pm 0.4}$ | 59.4$_{\pm 0.8}$ | 90.4$_{\pm 0.3}$ | 70.8 |
| GNorm | 82.5$_{\pm 0.3}$ | 75.8$_{\pm 1.3}$ | 62.9$_{\pm 0.5}$ | 55.3$_{\pm 1.0}$ | 95.2$_{\pm 0.1}$ | 74.3 |
| ReAct | 79.2$_{\pm 5.0}$ | 75.0$_{\pm 1.9}$ | 63.1$_{\pm 5.5}$ | 57.3$_{\pm 1.1}$ | 94.3$_{\pm 0.1}$ | 73.8 |
| VIM | 83.3$_{\pm 2.5}$ | 80.9$_{\pm 1.6}$ | 79.4$_{\pm 2.8}$ | 58.3$_{\pm 1.0}$ | 87.3$_{\pm 0.1}$ | 77.8 |
| MDS | 82.8$_{\pm 0.1}$ | 83.4$_{\pm 0.2}$ | 92.5$_{\pm 0.1}$ | 49.9$_{\pm 0.7}$ | 80.4$_{\pm 0.0}$ | 77.8 |
| RMDS | 59.4$_{\pm 0.1}$ | 72.8$_{\pm 0.4}$ | 60.9$_{\pm 0.1}$ | 62.3$_{\pm 0.9}$ | 89.0$_{\pm 0.0}$ | 68.9 |
| KNN | 96.2$_{\pm 0.0}$ | 84.1$_{\pm 0.3}$ | 95.0$_{\pm 0.5}$ | 61.8$_{\pm 0.9}$ | 90.7$_{\pm 0.0}$ | 85.6 |
| SHE | 95.0$_{\pm 0.1}$ | 80.6$_{\pm 0.4}$ | 84.8$_{\pm 0.1}$ | 45.0$_{\pm 1.0}$ | **95.3**$_{\pm 0.0}$ | 80.1 |
| **SPROD** | **98.9**$_{\pm 0.0}$ | **85.6**$_{\pm 0.6}$ | **98.1**$_{\pm 0.0}$ | **66.9**$_{\pm 0.7}$ | 94.4$_{\pm 0.0}$ | **88.8** |

Table 26: AUPR-IN and AUPR-OUT performance of all methods using ResNet-34 as the feature backbone.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 60.1$_{\pm 0.9}$ | 18.6$_{\pm 1.1}$ | 30.7$_{\pm 0.2}$ | 84.0$_{\pm 0.3}$ | 52.9$_{\pm 0.4}$ | 49.3 |
| Energy | 57.8$_{\pm 5.8}$ | 19.0$_{\pm 2.9}$ | 30.4$_{\pm 1.7}$ | 85.4$_{\pm 0.4}$ | 48.8$_{\pm 0.6}$ | 48.3 |
| MLS | 57.8$_{\pm 5.3}$ | 18.8$_{\pm 2.8}$ | 30.4$_{\pm 1.4}$ | 84.8$_{\pm 0.4}$ | 51.3$_{\pm 0.5}$ | 48.6 |
| KLM | 33.9$_{\pm 0.4}$ | 18.3$_{\pm 0.4}$ | 40.1$_{\pm 0.3}$ | 74.3$_{\pm 0.6}$ | 52.2$_{\pm 2.9}$ | 43.8 |
| GNorm | 69.5$_{\pm 0.3}$ | 14.5$_{\pm 0.3}$ | 36.8$_{\pm 0.3}$ | 78.5$_{\pm 0.3}$ | 51.7$_{\pm 0.2}$ | 50.2 |
| ReAct | 68.6$_{\pm 4.8}$ | 14.3$_{\pm 1.2}$ | 32.0$_{\pm 2.5}$ | 85.4$_{\pm 0.4}$ | 53.3$_{\pm 1.0}$ | 50.7 |
| VIM | 74.8$_{\pm 4.9}$ | 24.7$_{\pm 4.0}$ | 42.5$_{\pm 2.9}$ | 83.9$_{\pm 0.5}$ | 43.7$_{\pm 0.5}$ | 53.9 |
| MDS | 82.3$_{\pm 0.1}$ | 33.9$_{\pm 0.5}$ | 87.8$_{\pm 0.1}$ | 71.2$_{\pm 0.8}$ | 25.9$_{\pm 0.1}$ | 60.2 |
| RMDS | 41.1$_{\pm 0.1}$ | 14.4$_{\pm 0.3}$ | 31.7$_{\pm 0.1}$ | **86.2**$_{\pm 0.3}$ | 53.2$_{\pm 0.2}$ | 45.3 |
| KNN | 96.5$_{\pm 0.1}$ | 27.0$_{\pm 0.4}$ | 83.6$_{\pm 0.3}$ | 83.2$_{\pm 0.5}$ | 39.6$_{\pm 0.2}$ | 66.0 |
| SHE | 75.9$_{\pm 0.3}$ | 15.3$_{\pm 0.1}$ | 69.0$_{\pm 0.4}$ | 74.2$_{\pm 0.5}$ | 51.6$_{\pm 0.4}$ | 57.2 |
| **SPROD** | **97.0**$_{\pm 0.1}$ | **40.4**$_{\pm 0.6}$ | **94.5**$_{\pm 0.1}$ | 84.6$_{\pm 0.4}$ | **53.4**$_{\pm 0.2}$ | **74.0** |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 70.9$_{\pm 0.6}$ | 74.8$_{\pm 0.8}$ | 60.8$_{\pm 0.6}$ | 64.1$_{\pm 0.9}$ | 94.3$_{\pm 0.1}$ | 73.0 |
| Energy | 71.4$_{\pm 7.5}$ | 76.2$_{\pm 2.5}$ | 61.7$_{\pm 5.3}$ | 64.9$_{\pm 1.1}$ | 93.8$_{\pm 0.1}$ | 73.6 |
| MLS | 71.3$_{\pm 6.0}$ | 75.5$_{\pm 1.8}$ | 61.4$_{\pm 4.0}$ | 65.1$_{\pm 1.0}$ | 94.1$_{\pm 0.1}$ | 73.5 |
| KLM | 58.5$_{\pm 0.6}$ | 76.3$_{\pm 0.6}$ | 67.3$_{\pm 0.7}$ | 62.7$_{\pm 0.9}$ | 92.6$_{\pm 0.4}$ | 71.5 |
| GNorm | 79.0$_{\pm 0.5}$ | 73.0$_{\pm 0.6}$ | 63.6$_{\pm 0.5}$ | 62.6$_{\pm 1.0}$ | **95.8**$_{\pm 0.0}$ | 74.8 |
| ReAct | 75.5$_{\pm 6.7}$ | 74.1$_{\pm 3.9}$ | 61.3$_{\pm 5.4}$ | 64.4$_{\pm 1.1}$ | 95.2$_{\pm 0.2}$ | 74.1 |
| VIM | 85.7$_{\pm 3.3}$ | 80.2$_{\pm 1.1}$ | 78.7$_{\pm 2.3}$ | 60.8$_{\pm 1.0}$ | 89.6$_{\pm 0.1}$ | 79.0 |
| MDS | 88.8$_{\pm 0.0}$ | 84.0$_{\pm 0.5}$ | 92.2$_{\pm 0.1}$ | 49.7$_{\pm 0.8}$ | 83.6$_{\pm 0.0}$ | 79.7 |
| RMDS | 63.0$_{\pm 0.1}$ | 71.0$_{\pm 0.5}$ | 60.4$_{\pm 0.2}$ | 66.5$_{\pm 1.0}$ | 91.3$_{\pm 0.0}$ | 70.4 |
| KNN | 98.5$_{\pm 0.0}$ | 82.7$_{\pm 0.4}$ | 92.7$_{\pm 0.1}$ | 63.5$_{\pm 0.9}$ | 92.3$_{\pm 0.0}$ | 85.9 |
| SHE | 93.3$_{\pm 0.1}$ | 77.0$_{\pm 0.3}$ | 88.5$_{\pm 0.1}$ | 58.6$_{\pm 0.5}$ | 95.7$_{\pm 0.1}$ | 82.6 |
| **SPROD** | **98.9**$_{\pm 0.0}$ | **84.2**$_{\pm 0.7}$ | **96.3**$_{\pm 0.0}$ | **68.5**$_{\pm 0.9}$ | 95.4$_{\pm 0.0}$ | **88.7** |

Table 27: AUPR-IN and AUPR-OUT performance of all methods using ResNet-50 as the feature backbone.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 52.4$_{\pm 1.1}$ | 23.9$_{\pm 0.9}$ | 28.3$_{\pm 0.1}$ | 86.4$_{\pm 0.3}$ | 58.2$_{\pm 0.6}$ | 49.8 |
| Energy | 52.1$_{\pm 3.4}$ | 23.4$_{\pm 2.6}$ | 28.2$_{\pm 0.7}$ | 86.7$_{\pm 0.4}$ | 52.9$_{\pm 0.9}$ | 48.7 |
| MLS | 52.1$_{\pm 3.3}$ | 23.4$_{\pm 2.5}$ | 28.2$_{\pm 0.5}$ | 86.9$_{\pm 0.4}$ | 55.8$_{\pm 0.5}$ | 49.3 |
| KLM | 39.5$_{\pm 0.5}$ | 17.1$_{\pm 0.7}$ | 41.0$_{\pm 0.7}$ | 76.0$_{\pm 0.7}$ | 52.3$_{\pm 1.2}$ | 45.2 |
| GNorm | 70.2$_{\pm 0.5}$ | 15.8$_{\pm 0.2}$ | 34.5$_{\pm 0.3}$ | 78.4$_{\pm 0.5}$ | 56.3$_{\pm 0.3}$ | 51.0 |
| ReAct | 67.8$_{\pm 3.2}$ | 20.2$_{\pm 2.3}$ | 29.7$_{\pm 1.1}$ | 87.0$_{\pm 0.4}$ | **58.7**$_{\pm 1.5}$ | 52.7 |
| VIM | 72.7$_{\pm 3.6}$ | 27.3$_{\pm 3.0}$ | 39.2$_{\pm 1.2}$ | 86.5$_{\pm 0.4}$ | 51.5$_{\pm 1.2}$ | 55.4 |
| MDS | 88.0$_{\pm 0.1}$ | 30.2$_{\pm 0.5}$ | 88.4$_{\pm 0.1}$ | 72.6$_{\pm 0.7}$ | 28.9$_{\pm 0.1}$ | 61.6 |
| RMDS | 45.3$_{\pm 0.1}$ | 14.4$_{\pm 0.5}$ | 34.4$_{\pm 0.1}$ | 88.5$_{\pm 0.3}$ | 43.1$_{\pm 0.1}$ | 45.1 |
| KNN | 97.8$_{\pm 0.1}$ | 24.4$_{\pm 0.5}$ | 85.6$_{\pm 0.2}$ | 86.2$_{\pm 0.4}$ | 46.0$_{\pm 0.2}$ | 68.0 |
| SHE | 77.0$_{\pm 0.4}$ | 16.2$_{\pm 0.3}$ | 58.2$_{\pm 0.2}$ | 61.6$_{\pm 0.7}$ | 49.4$_{\pm 0.3}$ | 52.5 |
| **SPROD** | **98.1**$_{\pm 0.1}$ | **40.6**$_{\pm 1.2}$ | **96.7**$_{\pm 0.1}$ | **88.9**$_{\pm 0.4}$ | 58.3$_{\pm 0.2}$ | **76.5** |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 70.9$_{\pm 0.4}$ | 75.8$_{\pm 1.0}$ | 57.9$_{\pm 0.2}$ | 66.5$_{\pm 0.9}$ | 93.9$_{\pm 0.1}$ | 73.0 |
| Energy | 70.4$_{\pm 2.4}$ | 75.8$_{\pm 1.9}$ | 58.5$_{\pm 3.0}$ | 66.5$_{\pm 1.0}$ | 93.7$_{\pm 0.2}$ | 73.0 |
| MLS | 70.8$_{\pm 1.8}$ | 75.6$_{\pm 1.7}$ | 58.1$_{\pm 2.0}$ | 66.9$_{\pm 1.0}$ | 94.0$_{\pm 0.1}$ | 73.1 |
| KLM | 64.7$_{\pm 0.5}$ | 74.6$_{\pm 1.5}$ | 66.8$_{\pm 0.2}$ | 65.7$_{\pm 0.9}$ | 93.3$_{\pm 0.3}$ | 73.0 |
| GNorm | 81.5$_{\pm 0.5}$ | 73.7$_{\pm 0.9}$ | 60.7$_{\pm 0.3}$ | 63.1$_{\pm 0.9}$ | **96.1**$_{\pm 0.1}$ | 75.0 |
| ReAct | 76.2$_{\pm 4.6}$ | 77.7$_{\pm 3.5}$ | 60.0$_{\pm 3.6}$ | 66.5$_{\pm 1.0}$ | 95.5$_{\pm 0.2}$ | 75.2 |
| VIM | 85.9$_{\pm 1.6}$ | 79.5$_{\pm 1.3}$ | 78.7$_{\pm 1.0}$ | 64.2$_{\pm 1.0}$ | 92.4$_{\pm 0.3}$ | 80.1 |
| MDS | 92.5$_{\pm 0.0}$ | 82.8$_{\pm 0.3}$ | 94.6$_{\pm 0.0}$ | 51.2$_{\pm 0.9}$ | 85.6$_{\pm 0.0}$ | 81.3 |
| RMDS | 68.0$_{\pm 0.1}$ | 70.7$_{\pm 0.6}$ | 62.8$_{\pm 0.1}$ | 68.3$_{\pm 0.9}$ | 89.0$_{\pm 0.0}$ | 71.8 |
| KNN | 99.1$_{\pm 0.0}$ | 82.3$_{\pm 0.4}$ | 94.7$_{\pm 0.0}$ | 67.6$_{\pm 1.0}$ | 93.4$_{\pm 0.0}$ | 87.4 |
| SHE | 93.5$_{\pm 0.1}$ | 77.4$_{\pm 0.3}$ | 82.6$_{\pm 0.1}$ | 46.4$_{\pm 0.8}$ | 95.6$_{\pm 0.0}$ | 79.1 |
| **SPROD** | **99.2**$_{\pm 0.0}$ | **84.3**$_{\pm 0.6}$ | **98.1**$_{\pm 0.0}$ | **73.0**$_{\pm 0.9}$ | 96.0$_{\pm 0.0}$ | **90.1** |

Table 28: AUPR-IN and AUPR-OUT performance of all methods using ResNet-101 as the feature backbone.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 55.8$_{\pm 0.9}$ | 23.6$_{\pm 2.9}$ | 29.6$_{\pm 0.6}$ | 88.0$_{\pm 0.3}$ | 57.4$_{\pm 0.3}$ | 50.9 |
| Energy | 53.9$_{\pm 3.2}$ | 23.9$_{\pm 3.4}$ | 29.6$_{\pm 1.1}$ | 88.0$_{\pm 0.4}$ | 55.0$_{\pm 0.7}$ | 50.1 |
| MLS | 54.1$_{\pm 2.9}$ | 23.8$_{\pm 3.3}$ | 29.7$_{\pm 1.0}$ | 88.1$_{\pm 0.4}$ | 57.5$_{\pm 0.4}$ | 50.6 |
| KLM | 38.1$_{\pm 0.4}$ | 18.7$_{\pm 1.2}$ | 40.2$_{\pm 0.6}$ | 78.2$_{\pm 0.6}$ | 50.6$_{\pm 0.8}$ | 45.2 |
| GNorm | 64.2$_{\pm 0.5}$ | 19.0$_{\pm 0.6}$ | 37.5$_{\pm 0.7}$ | 78.8$_{\pm 0.4}$ | 57.1$_{\pm 0.2}$ | 51.3 |
| ReAct | 67.2$_{\pm 3.0}$ | 16.7$_{\pm 2.6}$ | 31.4$_{\pm 1.8}$ | 87.6$_{\pm 0.6}$ | 62.8$_{\pm 1.9}$ | 53.1 |
| VIM | 83.7$_{\pm 2.2}$ | 27.1$_{\pm 3.8}$ | 43.4$_{\pm 1.8}$ | 87.2$_{\pm 0.5}$ | 56.9$_{\pm 1.1}$ | 59.7 |
| MDS | 95.2$_{\pm 0.1}$ | 24.4$_{\pm 0.5}$ | 91.5$_{\pm 0.1}$ | 73.9$_{\pm 0.8}$ | 35.6$_{\pm 0.1}$ | 64.1 |
| RMDS | 44.8$_{\pm 0.0}$ | 13.8$_{\pm 0.2}$ | 31.6$_{\pm 0.1}$ | **89.8**$_{\pm 0.3}$ | 47.4$_{\pm 0.2}$ | 45.5 |
| KNN | 97.9$_{\pm 0.1}$ | 20.2$_{\pm 5.4}$ | 82.3$_{\pm 0.4}$ | 88.0$_{\pm 0.4}$ | 49.9$_{\pm 0.2}$ | 67.7 |
| SHE | 68.5$_{\pm 0.6}$ | 34.9$_{\pm 1.1}$ | 66.8$_{\pm 0.3}$ | 65.7$_{\pm 0.7}$ | 47.1$_{\pm 0.3}$ | 56.6 |
| **SPROD** | **98.1**$_{\pm 0.1}$ | **44.2**$_{\pm 1.2}$ | **97.3**$_{\pm 0.0}$ | 87.8$_{\pm 0.4}$ | **58.0**$_{\pm 0.2}$ | **77.1** |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 72.7$_{\pm 0.6}$ | 75.3$_{\pm 1.4}$ | 59.8$_{\pm 1.1}$ | 68.8$_{\pm 1.0}$ | 94.9$_{\pm 0.1}$ | 74.3 |
| Energy | 70.7$_{\pm 3.9}$ | 76.0$_{\pm 2.0}$ | 60.0$_{\pm 2.9}$ | 69.6$_{\pm 1.2}$ | 93.9$_{\pm 0.2}$ | 74.0 |
| MLS | 71.4$_{\pm 2.5}$ | 75.6$_{\pm 1.7}$ | 59.8$_{\pm 2.2}$ | 69.9$_{\pm 1.2}$ | 94.5$_{\pm 0.1}$ | 74.2 |
| KLM | 64.2$_{\pm 0.7}$ | 75.8$_{\pm 1.1}$ | 67.3$_{\pm 1.2}$ | 67.9$_{\pm 1.0}$ | 93.6$_{\pm 0.2}$ | 73.8 |
| GNorm | 79.8$_{\pm 0.6}$ | 76.1$_{\pm 1.1}$ | 63.0$_{\pm 1.0}$ | 65.6$_{\pm 1.1}$ | 96.0$_{\pm 0.1}$ | 76.1 |
| ReAct | 75.5$_{\pm 5.1}$ | 75.0$_{\pm 4.2}$ | 61.4$_{\pm 3.9}$ | 69.2$_{\pm 1.2}$ | 95.7$_{\pm 0.3}$ | 75.4 |
| VIM | 92.3$_{\pm 1.2}$ | 78.4$_{\pm 1.6}$ | 80.0$_{\pm 1.0}$ | 64.8$_{\pm 1.1}$ | 93.5$_{\pm 0.2}$ | 81.8 |
| MDS | 97.0$_{\pm 0.0}$ | 81.5$_{\pm 0.4}$ | 95.2$_{\pm 0.0}$ | 51.3$_{\pm 0.9}$ | 88.2$_{\pm 0.1}$ | 82.6 |
| RMDS | 69.9$_{\pm 0.1}$ | 69.8$_{\pm 0.2}$ | 60.2$_{\pm 0.1}$ | 71.2$_{\pm 1.0}$ | 90.8$_{\pm 0.0}$ | 72.4 |
| KNN | **99.4**$_{\pm 0.0}$ | 80.9$_{\pm 1.8}$ | 93.7$_{\pm 0.1}$ | 70.0$_{\pm 1.1}$ | 94.3$_{\pm 0.0}$ | 87.7 |
| SHE | 90.8$_{\pm 0.2}$ | **87.1**$_{\pm 0.5}$ | 88.2$_{\pm 0.1}$ | 51.6$_{\pm 1.2}$ | 95.1$_{\pm 0.1}$ | 82.6 |
| **SPROD** | **99.4**$_{\pm 0.0}$ | 84.8$_{\pm 0.7}$ | **98.5**$_{\pm 0.0}$ | **74.5**$_{\pm 0.9}$ | **96.5**$_{\pm 0.0}$ | **90.7** |

Table 29: AUPR-IN and AUPR-OUT performance of all methods using DINOv2-S as the feature backbone.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 75.8$_{\pm 2.4}$ | 14.8$_{\pm 1.0}$ | 54.0$_{\pm 0.6}$ | 92.1$_{\pm 0.3}$ | 60.0$_{\pm 1.1}$ | 59.3 |
| Energy | 71.7$_{\pm 3.3}$ | 15.9$_{\pm 2.8}$ | 53.1$_{\pm 9.3}$ | 93.4$_{\pm 0.2}$ | 65.6$_{\pm 1.5}$ | 59.9 |
| MLS | 72.2$_{\pm 3.0}$ | 15.7$_{\pm 2.7}$ | 53.3$_{\pm 8.8}$ | 93.4$_{\pm 0.2}$ | 61.4$_{\pm 1.5}$ | 59.2 |
| KLM | 64.7$_{\pm 1.6}$ | 23.5$_{\pm 1.9}$ | 47.2$_{\pm 6.2}$ | 84.4$_{\pm 0.8}$ | 60.8$_{\pm 1.6}$ | 56.1 |
| GNorm | 79.4$_{\pm 1.7}$ | 13.8$_{\pm 0.5}$ | 47.1$_{\pm 0.4}$ | 86.3$_{\pm 0.7}$ | **66.0**$_{\pm 1.3}$ | 58.5 |
| ReAct | 70.7$_{\pm 2.9}$ | 15.2$_{\pm 2.2}$ | 52.9$_{\pm 9.2}$ | 93.4$_{\pm 0.2}$ | 65.6$_{\pm 1.3}$ | 59.6 |
| VIM | 95.8$_{\pm 0.9}$ | 16.8$_{\pm 3.3}$ | 74.4$_{\pm 9.4}$ | 91.9$_{\pm 0.4}$ | 54.7$_{\pm 1.0}$ | 66.7 |
| MDS | 98.8$_{\pm 0.0}$ | 29.6$_{\pm 0.7}$ | 91.9$_{\pm 0.1}$ | 89.9$_{\pm 0.5}$ | 48.7$_{\pm 0.2}$ | 71.8 |
| RMDS | 51.2$_{\pm 0.1}$ | 13.0$_{\pm 0.3}$ | 35.5$_{\pm 0.2}$ | 93.9$_{\pm 0.3}$ | 57.5$_{\pm 0.3}$ | 50.2 |
| KNN | 98.9$_{\pm 0.1}$ | 16.0$_{\pm 0.2}$ | 68.5$_{\pm 1.2}$ | **94.2**$_{\pm 0.2}$ | 45.6$_{\pm 1.0}$ | 64.6 |
| SHE | 98.6$_{\pm 0.1}$ | 44.7$_{\pm 1.3}$ | 88.9$_{\pm 0.2}$ | 85.3$_{\pm 0.4}$ | 58.1$_{\pm 0.4}$ | 75.1 |
| **SPROD** | **99.4**$_{\pm 0.0}$ | **45.0**$_{\pm 1.1}$ | **95.3**$_{\pm 0.0}$ | 93.9$_{\pm 0.2}$ | 61.1$_{\pm 0.2}$ | **78.9** |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 82.3$_{\pm 1.6}$ | 71.9$_{\pm 1.3}$ | 72.8$_{\pm 0.3}$ | 77.1$_{\pm 0.8}$ | 95.5$_{\pm 0.2}$ | 79.9 |
| Energy | 81.1$_{\pm 3.2}$ | 73.7$_{\pm 4.3}$ | 73.3$_{\pm 8.4}$ | 78.0$_{\pm 1.1}$ | **96.1**$_{\pm 0.4}$ | 80.4 |
| MLS | 81.7$_{\pm 2.8}$ | 73.1$_{\pm 4.0}$ | 73.3$_{\pm 7.8}$ | 78.2$_{\pm 1.0}$ | 95.9$_{\pm 0.3}$ | 80.4 |
| KLM | 75.7$_{\pm 3.3}$ | 79.2$_{\pm 1.8}$ | 69.2$_{\pm 5.3}$ | 77.4$_{\pm 0.9}$ | 95.1$_{\pm 0.2}$ | 79.3 |
| GNorm | 83.8$_{\pm 1.4}$ | 71.5$_{\pm 1.1}$ | 72.7$_{\pm 0.3}$ | 73.6$_{\pm 1.1}$ | **96.1**$_{\pm 0.2}$ | 79.5 |
| ReAct | 80.5$_{\pm 3.0}$ | 73.3$_{\pm 4.0}$ | 73.1$_{\pm 8.1}$ | 78.0$_{\pm 1.1}$ | 96.0$_{\pm 0.4}$ | 80.2 |
| VIM | 98.5$_{\pm 0.3}$ | 75.1$_{\pm 3.6}$ | 88.9$_{\pm 4.8}$ | 74.7$_{\pm 1.2}$ | 95.9$_{\pm 0.1}$ | 86.6 |
| MDS | 99.3$_{\pm 0.0}$ | 81.2$_{\pm 0.3}$ | 95.3$_{\pm 0.0}$ | 72.3$_{\pm 1.3}$ | 95.6$_{\pm 0.0}$ | 88.7 |
| RMDS | 73.6$_{\pm 0.1}$ | 67.8$_{\pm 0.3}$ | 65.0$_{\pm 0.1}$ | 80.4$_{\pm 1.1}$ | 95.4$_{\pm 0.0}$ | 76.4 |
| KNN | 99.5$_{\pm 0.0}$ | 77.8$_{\pm 0.3}$ | 90.9$_{\pm 0.2}$ | 84.5$_{\pm 0.8}$ | 94.3$_{\pm 0.3}$ | 89.4 |
| SHE | 99.4$_{\pm 0.0}$ | 79.0$_{\pm 0.4}$ | 95.3$_{\pm 0.1}$ | 72.3$_{\pm 1.4}$ | 95.5$_{\pm 0.1}$ | 88.3 |
| **SPROD** | **99.7**$_{\pm 0.0}$ | **84.6**$_{\pm 0.4}$ | **97.1**$_{\pm 0.0}$ | **87.3**$_{\pm 0.5}$ | 96.0$_{\pm 0.0}$ | **92.9** |

Table 30: AUPR-IN and AUPR-OUT performance of all methods using ViT-S as the feature backbone.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 67.1$_{\pm 2.2}$ | 16.8$_{\pm 0.7}$ | 35.0$_{\pm 0.3}$ | 89.4$_{\pm 0.2}$ | 63.9$_{\pm 1.2}$ | 54.4 |
| Energy | 63.2$_{\pm 10.1}$ | 19.0$_{\pm 6.7}$ | 35.5$_{\pm 2.8}$ | 90.6$_{\pm 0.3}$ | 65.6$_{\pm 1.7}$ | 54.8 |
| MLS | 63.3$_{\pm 9.8}$ | 18.8$_{\pm 6.5}$ | 35.4$_{\pm 2.6}$ | 90.5$_{\pm 0.3}$ | 66.1$_{\pm 1.5}$ | 54.8 |
| KLM | 53.6$_{\pm 2.9}$ | 21.7$_{\pm 2.1}$ | 31.2$_{\pm 3.6}$ | 79.3$_{\pm 0.7}$ | 60.5$_{\pm 2.1}$ | 49.3 |
| GNorm | 73.1$_{\pm 1.4}$ | 16.4$_{\pm 0.5}$ | 34.6$_{\pm 0.5}$ | 87.8$_{\pm 0.4}$ | 65.6$_{\pm 0.6}$ | 55.5 |
| ReAct | 64.0$_{\pm 8.1}$ | 24.4$_{\pm 9.0}$ | 35.5$_{\pm 2.9}$ | 90.8$_{\pm 0.2}$ | **65.8**$_{\pm 2.0}$ | 56.1 |
| VIM | 81.0$_{\pm 8.8}$ | 21.8$_{\pm 8.7}$ | 51.7$_{\pm 5.8}$ | 91.3$_{\pm 0.3}$ | 64.6$_{\pm 1.3}$ | 62.1 |
| MDS | 94.4$_{\pm 0.1}$ | 38.1$_{\pm 1.2}$ | 93.8$_{\pm 0.1}$ | 87.1$_{\pm 0.4}$ | 58.9$_{\pm 0.2}$ | 74.5 |
| RMDS | 46.6$_{\pm 0.1}$ | 13.6$_{\pm 0.3}$ | 30.9$_{\pm 0.1}$ | **92.0**$_{\pm 0.2}$ | 64.5$_{\pm 0.2}$ | 49.5 |
| KNN | 92.8$_{\pm 0.1}$ | 30.6$_{\pm 6.9}$ | 89.3$_{\pm 0.1}$ | 88.3$_{\pm 0.4}$ | 52.8$_{\pm 0.2}$ | 70.8 |
| SHE | 92.1$_{\pm 0.3}$ | 38.9$_{\pm 1.0}$ | 87.8$_{\pm 0.3}$ | 80.4$_{\pm 0.7}$ | 58.5$_{\pm 0.2}$ | 71.5 |
| **SPROD** | **95.1**$_{\pm 0.1}$ | **42.0**$_{\pm 1.4}$ | **95.5**$_{\pm 0.1}$ | 88.4$_{\pm 0.3}$ | 64.4$_{\pm 0.3}$ | **77.1** |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 75.0$_{\pm 1.5}$ | 73.4$_{\pm 0.6}$ | 62.4$_{\pm 0.5}$ | 70.5$_{\pm 0.6}$ | 95.2$_{\pm 0.2}$ | 75.3 |
| Energy | 73.4$_{\pm 8.8}$ | 74.5$_{\pm 4.8}$ | 64.3$_{\pm 4.5}$ | 72.9$_{\pm 0.9}$ | 95.4$_{\pm 0.3}$ | 76.1 |
| MLS | 73.6$_{\pm 8.1}$ | 74.0$_{\pm 4.3}$ | 64.0$_{\pm 4.1}$ | 72.8$_{\pm 0.8}$ | 95.5$_{\pm 0.2}$ | 76.0 |
| KLM | 71.1$_{\pm 6.9}$ | 77.8$_{\pm 2.4}$ | 59.2$_{\pm 4.6}$ | 69.3$_{\pm 0.7}$ | 95.3$_{\pm 0.2}$ | 74.5 |
| GNorm | 77.5$_{\pm 1.3}$ | 73.2$_{\pm 0.6}$ | 62.5$_{\pm 0.5}$ | 70.8$_{\pm 0.8}$ | 95.9$_{\pm 0.1}$ | 76.0 |
| ReAct | 74.5$_{\pm 7.7}$ | 78.5$_{\pm 4.6}$ | 64.1$_{\pm 4.6}$ | 72.0$_{\pm 0.9}$ | 94.3$_{\pm 0.4}$ | 76.7 |
| VIM | 87.1$_{\pm 7.1}$ | 76.4$_{\pm 4.2}$ | 79.2$_{\pm 4.4}$ | 74.4$_{\pm 0.8}$ | 95.6$_{\pm 0.2}$ | 82.5 |
| MDS | 95.7$_{\pm 0.0}$ | 83.1$_{\pm 0.5}$ | 95.2$_{\pm 0.0}$ | 70.3$_{\pm 0.8}$ | 95.3$_{\pm 0.0}$ | 87.9 |
| RMDS | 69.3$_{\pm 0.1}$ | 68.8$_{\pm 0.4}$ | 59.1$_{\pm 0.1}$ | 76.7$_{\pm 0.7}$ | 95.9$_{\pm 0.0}$ | 74.0 |
| KNN | 94.4$_{\pm 0.0}$ | 80.8$_{\pm 1.2}$ | 93.4$_{\pm 0.0}$ | 72.1$_{\pm 0.8}$ | 93.6$_{\pm 0.0}$ | 86.9 |
| SHE | 95.3$_{\pm 0.2}$ | 77.7$_{\pm 0.3}$ | 93.8$_{\pm 0.2}$ | 70.1$_{\pm 0.9}$ | 95.0$_{\pm 0.1}$ | 86.4 |
| **SPROD** | **97.8**$_{\pm 0.0}$ | **83.3**$_{\pm 0.5}$ | **97.2**$_{\pm 0.0}$ | **77.0**$_{\pm 0.6}$ | **96.0**$_{\pm 0.0}$ | **90.3** |

Table 31: AUPR-IN and AUPR-OUT performance of all methods using Swin-B as the feature backbone.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 79.8$_{\pm 1.0}$ | 14.0$_{\pm 0.6}$ | 47.1$_{\pm 0.4}$ | 94.8$_{\pm 0.2}$ | 70.0$_{\pm 1.8}$ | 61.1 |
| Energy | 79.0$_{\pm 1.8}$ | 14.0$_{\pm 0.7}$ | 47.5$_{\pm 2.5}$ | 95.9$_{\pm 0.2}$ | 60.3$_{\pm 1.8}$ | 59.3 |
| MLS | 79.2$_{\pm 1.6}$ | 14.0$_{\pm 0.7}$ | 47.5$_{\pm 2.3}$ | 95.0$_{\pm 0.2}$ | 64.9$_{\pm 1.6}$ | 60.1 |
| KLM | 40.2$_{\pm 1.0}$ | 24.9$_{\pm 2.1}$ | 35.9$_{\pm 0.5}$ | 94.0$_{\pm 0.3}$ | 67.2$_{\pm 1.0}$ | 52.4 |
| GNorm | 43.1$_{\pm 0.8}$ | 12.7$_{\pm 0.2}$ | 35.1$_{\pm 0.2}$ | 83.8$_{\pm 1.0}$ | 19.2$_{\pm 0.3}$ | 38.8 |
| ReAct | 80.9$_{\pm 1.7}$ | 15.2$_{\pm 0.8}$ | 48.5$_{\pm 2.3}$ | **96.1**$_{\pm 0.2}$ | 74.5$_{\pm 1.4}$ | 63.0 |
| VIM | 89.5$_{\pm 0.4}$ | 18.3$_{\pm 1.6}$ | 65.9$_{\pm 2.1}$ | 93.6$_{\pm 0.2}$ | 77.0$_{\pm 0.4}$ | 68.9 |
| MDS | 83.7$_{\pm 0.1}$ | 34.4$_{\pm 0.8}$ | 79.9$_{\pm 0.3}$ | 85.3$_{\pm 0.5}$ | 76.6$_{\pm 0.2}$ | 72.0 |
| RMDS | 72.7$_{\pm 0.1}$ | 13.7$_{\pm 0.3}$ | 35.4$_{\pm 0.1}$ | 96.4$_{\pm 0.1}$ | 76.6$_{\pm 0.2}$ | 59.0 |
| KNN | 86.8$_{\pm 0.1}$ | 29.8$_{\pm 0.8}$ | 88.2$_{\pm 0.2}$ | 95.2$_{\pm 0.2}$ | 73.0$_{\pm 0.3}$ | 74.6 |
| SHE | 96.6$_{\pm 0.1}$ | 27.2$_{\pm 0.9}$ | 95.4$_{\pm 0.2}$ | 95.9$_{\pm 0.2}$ | 59.7$_{\pm 0.2}$ | 75.0 |
| **SPROD** | **98.3**$_{\pm 0.1}$ | **47.8**$_{\pm 1.4}$ | **98.3**$_{\pm 0.0}$ | 95.2$_{\pm 0.2}$ | **80.9**$_{\pm 0.2}$ | **84.1** |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 84.5$_{\pm 0.9}$ | 69.7$_{\pm 0.8}$ | 69.0$_{\pm 0.4}$ | 86.2$_{\pm 0.5}$ | 97.3$_{\pm 0.1}$ | 81.3 |
| Energy | 84.4$_{\pm 1.9}$ | 70.0$_{\pm 1.3}$ | 69.5$_{\pm 2.9}$ | **87.4**$_{\pm 0.6}$ | 96.1$_{\pm 0.2}$ | 81.5 |
| MLS | 84.6$_{\pm 1.4}$ | 69.8$_{\pm 1.2}$ | 69.3$_{\pm 2.2}$ | 87.1$_{\pm 0.6}$ | 97.0$_{\pm 0.1}$ | 81.6 |
| KLM | 72.6$_{\pm 1.3}$ | 77.4$_{\pm 1.8}$ | 65.6$_{\pm 0.5}$ | 85.2$_{\pm 0.4}$ | 97.0$_{\pm 0.2}$ | 79.6 |
| GNorm | 76.3$_{\pm 0.9}$ | 69.2$_{\pm 0.6}$ | 67.3$_{\pm 0.4}$ | 81.3$_{\pm 0.9}$ | 86.6$_{\pm 0.5}$ | 76.1 |
| ReAct | 85.2$_{\pm 2.1}$ | 71.4$_{\pm 1.4}$ | 69.8$_{\pm 2.7}$ | 87.0$_{\pm 0.6}$ | 97.0$_{\pm 0.2}$ | 82.1 |
| VIM | 91.7$_{\pm 0.2}$ | 74.8$_{\pm 1.6}$ | 86.0$_{\pm 0.9}$ | 73.5$_{\pm 0.8}$ | 97.9$_{\pm 0.0}$ | 84.8 |
| MDS | 88.1$_{\pm 0.1}$ | 84.2$_{\pm 0.7}$ | 91.1$_{\pm 0.1}$ | 58.3$_{\pm 1.1}$ | 97.8$_{\pm 0.0}$ | 83.9 |
| RMDS | 86.5$_{\pm 0.0}$ | 69.0$_{\pm 0.5}$ | 63.9$_{\pm 0.1}$ | 81.8$_{\pm 0.5}$ | 98.0$_{\pm 0.0}$ | 79.8 |
| KNN | 93.3$_{\pm 0.0}$ | 82.5$_{\pm 0.5}$ | 94.8$_{\pm 0.1}$ | 85.5$_{\pm 0.6}$ | 97.4$_{\pm 0.0}$ | 90.7 |
| SHE | 98.0$_{\pm 0.0}$ | 77.2$_{\pm 0.5}$ | 97.5$_{\pm 0.1}$ | 83.8$_{\pm 0.8}$ | 96.8$_{\pm 0.0}$ | 90.7 |
| **SPROD** | **99.3**$_{\pm 0.0}$ | **86.3**$_{\pm 0.6}$ | **99.2**$_{\pm 0.0}$ | 86.6$_{\pm 0.4}$ | **98.1**$_{\pm 0.0}$ | **93.9** |

Table 32: AUPR-IN and AUPR-OUT performance of all methods using DeiT-B as the feature backbone.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 72.9$_{\pm 1.2}$ | 21.4$_{\pm 1.4}$ | 40.5$_{\pm 0.2}$ | 92.4$_{\pm 0.3}$ | 67.9$_{\pm 1.1}$ | 59.0 |
| Energy | 73.7$_{\pm 3.4}$ | 22.2$_{\pm 3.5}$ | 39.8$_{\pm 2.4}$ | 92.4$_{\pm 0.5}$ | 65.6$_{\pm 1.8}$ | 58.7 |
| MLS | 73.6$_{\pm 3.2}$ | 22.2$_{\pm 3.4}$ | 39.9$_{\pm 2.2}$ | 92.4$_{\pm 0.5}$ | 67.1$_{\pm 1.1}$ | 59.0 |
| KLM | 35.3$_{\pm 0.3}$ | 18.6$_{\pm 0.6}$ | 35.7$_{\pm 0.4}$ | 87.5$_{\pm 0.5}$ | 65.1$_{\pm 1.7}$ | 48.4 |
| GNorm | 51.7$_{\pm 1.1}$ | 17.1$_{\pm 0.5}$ | 37.1$_{\pm 0.1}$ | 77.9$_{\pm 1.1}$ | 30.9$_{\pm 2.4}$ | 42.9 |
| ReAct | 73.6$_{\pm 3.4}$ | 20.3$_{\pm 2.9}$ | 40.3$_{\pm 2.6}$ | **93.1**$_{\pm 0.4}$ | 69.4$_{\pm 1.7}$ | 59.3 |
| VIM | 81.7$_{\pm 1.9}$ | 23.5$_{\pm 3.4}$ | 52.1$_{\pm 2.7}$ | 90.1$_{\pm 0.3}$ | 68.0$_{\pm 0.7}$ | 63.1 |
| MDS | 69.6$_{\pm 0.2}$ | 19.4$_{\pm 0.1}$ | 68.7$_{\pm 0.3}$ | 79.1$_{\pm 0.6}$ | 67.0$_{\pm 0.3}$ | 60.8 |
| RMDS | 56.3$_{\pm 0.1}$ | 13.9$_{\pm 0.2}$ | 36.2$_{\pm 0.3}$ | 92.8$_{\pm 0.2}$ | 69.3$_{\pm 0.3}$ | 53.7 |
| KNN | 88.6$_{\pm 0.8}$ | 18.5$_{\pm 2.6}$ | 82.9$_{\pm 0.2}$ | 89.5$_{\pm 0.4}$ | 66.7$_{\pm 0.3}$ | 69.2 |
| SHE | **93.8**$_{\pm 0.0}$ | **35.3**$_{\pm 0.6}$ | 92.4$_{\pm 0.1}$ | 80.6$_{\pm 1.0}$ | 65.4$_{\pm 0.4}$ | 73.5 |
| SPROD | 93.6$_{\pm 0.1}$ | 32.1$_{\pm 0.7}$ | **95.2**$_{\pm 0.0}$ | 92.1$_{\pm 0.2}$ | **69.7**$_{\pm 0.3}$ | **76.5** |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 77.9$_{\pm 0.7}$ | 74.7$_{\pm 1.0}$ | 66.1$_{\pm 0.3}$ | 80.8$_{\pm 0.7}$ | 94.9$_{\pm 0.2}$ | 78.9 |
| Energy | 80.1$_{\pm 3.6}$ | 75.6$_{\pm 2.2}$ | 65.7$_{\pm 3.3}$ | **83.5**$_{\pm 0.9}$ | 94.9$_{\pm 0.5}$ | 80.0 |
| MLS | 79.9$_{\pm 2.8}$ | 75.2$_{\pm 1.8}$ | 65.8$_{\pm 2.6}$ | 82.9$_{\pm 0.8}$ | 95.1$_{\pm 0.2}$ | 79.8 |
| KLM | 64.3$_{\pm 0.5}$ | 74.7$_{\pm 1.2}$ | 64.0$_{\pm 0.2}$ | 78.3$_{\pm 0.6}$ | 95.8$_{\pm 0.3}$ | 75.4 |
| GNorm | 75.5$_{\pm 0.6}$ | 75.6$_{\pm 0.8}$ | 66.6$_{\pm 0.3}$ | 75.1$_{\pm 1.2}$ | 90.3$_{\pm 1.2}$ | 76.6 |
| ReAct | 80.1$_{\pm 3.4}$ | 74.7$_{\pm 2.2}$ | 66.1$_{\pm 3.6}$ | **83.5**$_{\pm 0.9}$ | 95.4$_{\pm 0.5}$ | 80.0 |
| VIM | 84.2$_{\pm 1.5}$ | 77.9$_{\pm 1.5}$ | 77.3$_{\pm 2.0}$ | 66.1$_{\pm 0.9}$ | 95.7$_{\pm 0.1}$ | 80.2 |
| MDS | 79.8$_{\pm 0.1}$ | 80.8$_{\pm 0.4}$ | 84.4$_{\pm 0.1}$ | 50.8$_{\pm 1.1}$ | 95.7$_{\pm 0.0}$ | 78.3 |
| RMDS | 75.5$_{\pm 0.1}$ | 70.1$_{\pm 0.1}$ | 64.1$_{\pm 0.2}$ | 72.3$_{\pm 0.6}$ | 96.6$_{\pm 0.0}$ | 75.7 |
| KNN | 91.4$_{\pm 0.5}$ | 79.3$_{\pm 1.4}$ | 91.3$_{\pm 0.0}$ | 72.0$_{\pm 1.0}$ | 95.1$_{\pm 0.0}$ | 85.8 |
| SHE | 94.3$_{\pm 0.0}$ | 79.3$_{\pm 0.4}$ | 96.0$_{\pm 0.0}$ | 57.6$_{\pm 1.1}$ | 95.7$_{\pm 0.0}$ | 84.6 |
| SPROD | **94.8**$_{\pm 0.0}$ | **81.9**$_{\pm 0.5}$ | **97.0**$_{\pm 0.0}$ | 83.3$_{\pm 0.6}$ | **96.7**$_{\pm 0.0}$ | **90.7** |

Table 33: AUPR-IN and AUPR-OUT performance of all methods using ConvNeXt-B as the feature backbone.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 85.4$_{\pm 1.0}$ | 14.1$_{\pm 1.1}$ | 48.5$_{\pm 0.6}$ | 94.5$_{\pm 0.1}$ | 78.2$_{\pm 0.8}$ | 64.1 |
| Energy | 82.7$_{\pm 4.8}$ | 13.5$_{\pm 1.1}$ | 46.0$_{\pm 4.1}$ | 96.1$_{\pm 0.1}$ | 77.3$_{\pm 1.3}$ | 63.1 |
| MLS | 83.1$_{\pm 4.7}$ | 13.4$_{\pm 1.0}$ | 46.1$_{\pm 3.8}$ | 95.1$_{\pm 0.1}$ | 77.5$_{\pm 1.3}$ | 63.0 |
| KLM | 44.3$_{\pm 1.2}$ | 23.9$_{\pm 1.3}$ | 35.6$_{\pm 0.5}$ | 94.1$_{\pm 0.3}$ | 66.2$_{\pm 0.8}$ | 52.8 |
| GNorm | 51.0$_{\pm 1.4}$ | 21.0$_{\pm 1.1}$ | 31.1$_{\pm 0.2}$ | 87.8$_{\pm 0.6}$ | 66.4$_{\pm 1.4}$ | 51.5 |
| ReAct | 82.1$_{\pm 5.4}$ | 12.8$_{\pm 0.7}$ | 46.1$_{\pm 4.1}$ | 96.1$_{\pm 0.1}$ | 75.8$_{\pm 1.3}$ | 62.6 |
| VIM | 96.6$_{\pm 0.6}$ | 14.4$_{\pm 1.6}$ | 73.0$_{\pm 4.6}$ | **96.8**$_{\pm 0.1}$ | 81.5$_{\pm 0.5}$ | 72.5 |
| MDS | 97.0$_{\pm 0.1}$ | 25.0$_{\pm 0.8}$ | 94.9$_{\pm 0.1}$ | 92.6$_{\pm 0.3}$ | 75.3$_{\pm 0.2}$ | 77.0 |
| RMDS | 73.2$_{\pm 0.1}$ | 13.0$_{\pm 0.3}$ | 34.4$_{\pm 0.2}$ | **96.8**$_{\pm 0.1}$ | 78.6$_{\pm 0.2}$ | 59.2 |
| KNN | 94.3$_{\pm 0.1}$ | 28.3$_{\pm 0.3}$ | 85.7$_{\pm 0.2}$ | 94.9$_{\pm 0.2}$ | 79.3$_{\pm 0.2}$ | 76.5 |
| SHE | 92.8$_{\pm 0.1}$ | **49.7**$_{\pm 1.0}$ | 98.7$_{\pm 0.1}$ | 83.5$_{\pm 0.7}$ | 75.6$_{\pm 0.1}$ | 80.1 |
| SPROD | **98.4**$_{\pm 0.1}$ | 45.0$_{\pm 0.9}$ | **99.6**$_{\pm 0.0}$ | 95.3$_{\pm 0.6}$ | **83.2**$_{\pm 0.2}$ | **84.3** |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 89.6$_{\pm 0.6}$ | 69.5$_{\pm 0.9}$ | 68.6$_{\pm 0.5}$ | 84.0$_{\pm 0.5}$ | 97.7$_{\pm 0.1}$ | 81.9 |
| Energy | 88.4$_{\pm 3.7}$ | 69.5$_{\pm 2.2}$ | 67.9$_{\pm 6.1}$ | 85.7$_{\pm 0.6}$ | 97.9$_{\pm 0.1}$ | 81.9 |
| MLS | 88.9$_{\pm 3.2}$ | 69.3$_{\pm 1.9}$ | 68.0$_{\pm 5.4}$ | 85.0$_{\pm 0.6}$ | 98.0$_{\pm 0.1}$ | 81.8 |
| KLM | 80.0$_{\pm 1.1}$ | 75.0$_{\pm 1.5}$ | 65.4$_{\pm 0.3}$ | 87.6$_{\pm 0.4}$ | 96.7$_{\pm 0.2}$ | 80.9 |
| GNorm | 82.1$_{\pm 0.6}$ | 75.0$_{\pm 0.8}$ | 64.0$_{\pm 0.5}$ | 78.1$_{\pm 0.9}$ | 96.9$_{\pm 0.1}$ | 79.2 |
| ReAct | 87.7$_{\pm 3.5}$ | 68.3$_{\pm 1.8}$ | 67.9$_{\pm 5.9}$ | 85.1$_{\pm 0.6}$ | 97.9$_{\pm 0.1}$ | 81.4 |
| VIM | 98.1$_{\pm 0.4}$ | 71.5$_{\pm 2.0}$ | 89.8$_{\pm 2.0}$ | **88.9**$_{\pm 0.5}$ | 97.6$_{\pm 0.1}$ | 89.2 |
| MDS | 98.1$_{\pm 0.0}$ | 80.5$_{\pm 0.7}$ | 96.7$_{\pm 0.0}$ | 79.2$_{\pm 0.7}$ | 96.8$_{\pm 0.0}$ | 90.3 |
| RMDS | 89.6$_{\pm 0.0}$ | 67.9$_{\pm 0.3}$ | 63.9$_{\pm 0.2}$ | 84.4$_{\pm 0.5}$ | 97.7$_{\pm 0.0}$ | 80.7 |
| KNN | 97.8$_{\pm 0.0}$ | 82.7$_{\pm 0.6}$ | 96.6$_{\pm 0.1}$ | 80.5$_{\pm 0.7}$ | 98.1$_{\pm 0.0}$ | 91.1 |
| SHE | 96.8$_{\pm 0.1}$ | 82.8$_{\pm 0.7}$ | 99.5$_{\pm 0.0}$ | 61.1$_{\pm 0.9}$ | 97.7$_{\pm 0.1}$ | 87.6 |
| SPROD | **99.4**$_{\pm 0.0}$ | **85.3**$_{\pm 0.7}$ | **99.8**$_{\pm 0.0}$ | 86.0$_{\pm 0.7}$ | **98.4**$_{\pm 0.0}$ | **93.8** |

Table 34: AUPR-IN and AUPR-OUT performance of all methods using BiT-R50x1 as the feature backbone.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 59.3$_{\pm 3.0}$ | 18.8$_{\pm 3.9}$ | 33.4$_{\pm 0.8}$ | 89.8$_{\pm 0.2}$ | 55.4$_{\pm 0.6}$ | 51.3 |
| Energy | 56.4$_{\pm 2.9}$ | 20.5$_{\pm 7.5}$ | 32.0$_{\pm 3.7}$ | 91.3$_{\pm 0.3}$ | 52.0$_{\pm 0.8}$ | 50.4 |
| MLS | 56.5$_{\pm 2.9}$ | 20.5$_{\pm 7.5}$ | 32.0$_{\pm 3.5}$ | 91.3$_{\pm 0.3}$ | 52.7$_{\pm 0.7}$ | 50.6 |
| KLM | 35.7$_{\pm 0.5}$ | 18.9$_{\pm 2.0}$ | 36.8$_{\pm 0.6}$ | 79.7$_{\pm 0.6}$ | 56.5$_{\pm 1.0}$ | 45.5 |
| GNorm | 38.2$_{\pm 0.8}$ | 14.8$_{\pm 0.8}$ | 55.0$_{\pm 1.6}$ | 74.1$_{\pm 0.7}$ | 20.5$_{\pm 0.7}$ | 40.5 |
| ReAct | 47.6$_{\pm 2.9}$ | 14.9$_{\pm 2.1}$ | 33.7$_{\pm 4.2}$ | 86.9$_{\pm 0.4}$ | 55.9$_{\pm 0.9}$ | 47.8 |
| VIM | 90.7$_{\pm 2.3}$ | 25.7$_{\pm 9.9}$ | 55.3$_{\pm 9.8}$ | **92.3**$_{\pm 0.2}$ | **58.1**$_{\pm 0.4}$ | 64.4 |
| MDS | **98.8**$_{\pm 0.1}$ | 41.0$_{\pm 1.1}$ | **98.3**$_{\pm 0.0}$ | 85.5$_{\pm 0.6}$ | 52.7$_{\pm 0.3}$ | 75.3 |
| RMDS | 47.2$_{\pm 0.1}$ | 13.5$_{\pm 0.6}$ | 35.4$_{\pm 0.2}$ | 91.2$_{\pm 0.2}$ | 56.6$_{\pm 0.3}$ | 48.8 |
| KNN | 97.4$_{\pm 0.1}$ | 20.9$_{\pm 8.1}$ | 93.1$_{\pm 0.1}$ | 89.4$_{\pm 0.4}$ | 48.2$_{\pm 0.2}$ | 69.8 |
| SHE | 75.5$_{\pm 0.8}$ | 24.4$_{\pm 1.0}$ | 95.8$_{\pm 0.1}$ | 79.6$_{\pm 0.5}$ | 30.7$_{\pm 0.2}$ | 61.2 |
| **SPROD** | 97.3$_{\pm 0.1}$ | **49.8**$_{\pm 1.1}$ | **98.3**$_{\pm 0.0}$ | 89.9$_{\pm 0.3}$ | 56.1$_{\pm 0.3}$ | **78.3** |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| MSP | 76.0$_{\pm 1.1}$ | 74.7$_{\pm 2.3}$ | 62.6$_{\pm 0.8}$ | 73.5$_{\pm 0.6}$ | 94.8$_{\pm 0.2}$ | 76.3 |
| Energy | 72.8$_{\pm 2.5}$ | 75.7$_{\pm 4.1}$ | 60.8$_{\pm 6.4}$ | 79.5$_{\pm 0.8}$ | 94.3$_{\pm 0.3}$ | 76.6 |
| MLS | 73.3$_{\pm 2.3}$ | 75.5$_{\pm 4.0}$ | 60.9$_{\pm 5.7}$ | 78.7$_{\pm 0.7}$ | 94.6$_{\pm 0.3}$ | 76.6 |
| KLM | 65.3$_{\pm 0.5}$ | 74.2$_{\pm 3.1}$ | 64.5$_{\pm 0.9}$ | 71.0$_{\pm 0.7}$ | 94.9$_{\pm 0.2}$ | 74.0 |
| GNorm | 65.6$_{\pm 1.1}$ | 72.9$_{\pm 1.7}$ | 68.1$_{\pm 1.0}$ | 67.6$_{\pm 0.8}$ | 84.6$_{\pm 0.8}$ | 71.8 |
| ReAct | 64.7$_{\pm 2.0}$ | 74.2$_{\pm 4.6}$ | 61.7$_{\pm 6.8}$ | 72.6$_{\pm 0.9}$ | 94.2$_{\pm 0.5}$ | 73.5 |
| VIM | 95.8$_{\pm 1.0}$ | 78.0$_{\pm 4.3}$ | 84.8$_{\pm 4.6}$ | 75.4$_{\pm 0.8}$ | **96.5**$_{\pm 0.1}$ | 86.1 |
| MDS | **99.4**$_{\pm 0.0}$ | 83.9$_{\pm 0.4}$ | 98.8$_{\pm 0.0}$ | 65.1$_{\pm 1.0}$ | 95.5$_{\pm 0.0}$ | 88.5 |
| RMDS | 73.4$_{\pm 0.1}$ | 69.1$_{\pm 1.1}$ | 64.7$_{\pm 0.2}$ | 71.4$_{\pm 0.7}$ | 95.8$_{\pm 0.0}$ | 74.9 |
| KNN | 99.0$_{\pm 0.0}$ | 80.3$_{\pm 0.8}$ | 97.2$_{\pm 0.0}$ | 76.9$_{\pm 0.8}$ | 93.9$_{\pm 0.0}$ | 89.5 |
| SHE | 81.9$_{\pm 0.6}$ | 76.7$_{\pm 0.3}$ | 98.2$_{\pm 0.0}$ | 67.6$_{\pm 0.9}$ | 87.6$_{\pm 0.2}$ | 82.4 |
| **SPROD** | 99.1$_{\pm 0.0}$ | **85.5**$_{\pm 0.5}$ | **99.0**$_{\pm 0.0}$ | **83.8**$_{\pm 0.5}$ | 95.9$_{\pm 0.0}$ | **92.7** |

Table 35: Comparison of SPROD variants on SP-OOD datasets using AUROC and FPR@95 metrics. All methods utilize a pretrained ResNet-50 backbone. Values are averaged over five runs.

**AUROC↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| SPROD-Default | **98.8**$_{\pm 0.0}$ | **61.6**$_{\pm 0.9}$ | **97.4**$_{\pm 0.0}$ | 82.1$_{\pm 0.0}$ | 85.3$_{\pm 0.0}$ | **85.0** |
| SPROD-Converged | 98.5$_{\pm 0.0}$ | 59.1$_{\pm 0.9}$ | 97.1$_{\pm 0.1}$ | **83.0**$_{\pm 0.5}$ | 85.3$_{\pm 0.0}$ | 84.6 |
| SPROD-KMeans | 98.3$_{\pm 0.1}$ | 57.3$_{\pm 2.2}$ | 96.8$_{\pm 0.1}$ | 82.8$_{\pm 0.5}$ | **85.9**$_{\pm 0.0}$ | 84.2 |

**FPR@95↓**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| SPROD-Default | **4.7**$_{\pm 0.1}$ | **93.7**$_{\pm 0.9}$ | **19.0**$_{\pm 0.4}$ | **70.1**$_{\pm 0.0}$ | 58.0$_{\pm 0.1}$ | **49.1** |
| SPROD-Converged | 6.4$_{\pm 0.1}$ | 93.9$_{\pm 1.0}$ | 19.1$_{\pm 0.4}$ | 71.0$_{\pm 1.2}$ | 58.1$_{\pm 0.1}$ | 49.7 |
| SPROD-KMeans | 6.0$_{\pm 0.5}$ | **93.7**$_{\pm 1.2}$ | 19.3$_{\pm 0.4}$ | 71.3$_{\pm 1.1}$ | **56.4**$_{\pm 0.1}$ | 49.3 |

Table 36: Comparison of SPROD variants on SP-OOD datasets using AUPR-IN and AUPR-OUT metrics. All methods utilize a pretrained ResNet-50 backbone. Values are averaged over five runs.

**AUPR-IN↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| SPROD-Default | **98.1**$_{\pm 0.1}$ | **40.6**$_{\pm 1.2}$ | **96.7**$_{\pm 0.1}$ | **88.9**$_{\pm 0.0}$ | 58.3$_{\pm 0.0}$ | **76.5** |
| SPROD-Converged | 97.5$_{\pm 0.1}$ | 36.5$_{\pm 1.2}$ | 96.0$_{\pm 0.1}$ | 88.2$_{\pm 0.4}$ | 55.8$_{\pm 0.2}$ | 74.8 |
| SPROD-KMeans | 97.6$_{\pm 0.1}$ | 32.1$_{\pm 4.2}$ | 96.0$_{\pm 0.1}$ | 88.4$_{\pm 0.4}$ | **58.6**$_{\pm 0.3}$ | 74.5 |

**AUPR-OUT↑**

| Method | WB | CA | UC | AMC | SpI | Avg. |
|---|---|---|---|---|---|---|
| SPROD-Default | **99.2**$_{\pm 0.0}$ | **84.3**$_{\pm 0.6}$ | **98.1**$_{\pm 0.0}$ | 73.0$_{\pm 0.8}$ | 96.0$_{\pm 0.0}$ | **90.1** |
| SPROD-Converged | 99.1$_{\pm 0.0}$ | 83.6$_{\pm 0.6}$ | 98.0$_{\pm 0.0}$ | 73.2$_{\pm 0.8}$ | 96.0$_{\pm 0.0}$ | 90.0 |
| SPROD-KMeans | 99.1$_{\pm 0.0}$ | 83.4$_{\pm 0.6}$ | 98.0$_{\pm 0.0}$ | **73.4**$_{\pm 0.8}$ | **96.3**$_{\pm 0.0}$ | 90.0 |

## NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contributions, which include: (1) the proposal of SPROD, a novel post-hoc prototype-based OOD detection method designed to be robust against unknown spurious correlations; (2) the method's ability to operate without additional OOD data, hyperparameter tuning, or group annotations; (3) its broad applicability and superior performance, demonstrated through comprehensive experiments on diverse SP-OOD datasets, including the newly introduced AnimalsMetaCoCo dataset; and (4) insights into factors like backbone fine-tuning and scoring mechanisms in the context of SP-OOD. These claims are consistently addressed and supported by the methodology (Section 4), experimental results (Section 5.3, Table 2, Figures 4-6), and dataset descriptions (Section 5.2).

   Guidelines:

   - The answer NA 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 NA 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: Section 6, discusses limitations.

   Guidelines:

   - The answer NA 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: [NA]

   Justification: The paper proposes an algorithmic method (SPROD) and provides extensive empirical validation. It does not present new theoretical results in the form of theorems or formal proofs. The motivation in Section 3.2 discusses probabilistic concepts but does not derive new theoretical claims. Section 6 suggests that theoretical justifications are a direction for future work.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The paper provides substantial information to facilitate the reproduction of its main experimental results. 1. The proposed SPROD algorithm is described in detail in Section 4 and 1, with its three stages (Initial Prototype Construction, Classification-Aware Prototype Calculation, and Group Prototype Refinement) clearly outlined. Algorithm 1 provides pseudocode for SPROD. 2. Section 5.1 details the 11 baseline methods used for comparison, the evaluation metrics (AUROC, FPR@95), the primary backbone (ResNet-50, with ResNet-18 also mentioned for some analyses), and the fact that experiments were repeated five times with different random seeds for statistical robustness. 3. Section 5.2 thoroughly describes the five SP-OOD datasets used, including references to established ones (Waterbirds, CelebA, UrbanCars, Spurious ImageNet) and a description of the newly introduced AnimalsMetaCoCo dataset, clarifying the SP-OOD setup for each. Table 1 summarizes these datasets. 4. A key aspect aiding reproducibility is that SPROD is presented as a hyperparameter-free method. 5. The authors state their intention to release the code. The implementation is based on the OpenOOD codebase, which is open-source, and the code for SPROD and the experiments will be made publicly available. This will significantly aid in direct replication of the reported results.

   Guidelines:

   - The answer NA 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: We will make the code publicly available. The implementation is based on the open-source OpenOOD codebase, and the new method (SPROD) and experimental scripts will be added to this and released. For an anonymized version of the code refer to the supplementary material. The datasets used are: 1. Waterbirds [33], CelebA [34], UrbanCars [35], and Spurious ImageNet [36] are existing publicly available benchmarks, with references provided for access. 2. AnimalsMetaCoCo is a new benchmark dataset introduced in this paper, curated from MetaCoCo [71] (which is publicly available). The paper describes its construction 5.2, and its release is implied as part of the contributions and benchmarking efforts. Sufficient instructions for reproducing results are provided in the supplemental material and will accompany the code release. The paper itself (Sections 4, 5.1, 5.2, and Algorithm 1) details the method and experimental setup.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.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://nips.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, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the training and test details necessary to understand and reproduce the results. It outlines the datasets used (Section 5.2), their spurious correlation structure, evaluation metrics (e.g., AUROC, FPR@95), and competing baselines. The main experimental Setup is in Section 5.1, and further details are provided in the Supplemental.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We repeat all experiments five times with different random seeds and report the mean and standard deviation, as stated in Section 5.1. Experimental results reported primarily in Section 5.3 and supplemental.

   Guidelines:

   - The answer NA 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: Experiments were designed to be post-hoc and computationally efficient. All experiments were conducted on a single GeForce RTX 3090 Ti GPU, demonstrating the method's practicality in terms of resource requirements. Detailed information on compute resources, including execution setup and runtime considerations, is provided in the supplementary material.

Guidelines:

- The answer NA 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: Research presented in this paper fully conforms to the NeurIPS Code of Ethics. All datasets used are publicly available and widely adopted in the community. The methodology does not involve human subjects, personally identifiable information, or any potentially harmful applications. Efforts were made to ensure transparency, reproducibility, and fairness throughout the study.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: The Paper highlights positive societal impacts such as improving reliability and fairness in safety-critical applications in the introduction. While no direct negative impacts are identified, responsible deployment is encouraged. Additional discussion is provided in the supplementary material.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA 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., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not involve releasing models or datasets with high risk for misuse. It builds on publicly available datasets and does not introduce generative models or large pretrained components requiring additional safeguards.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: The paper uses publicly available datasets and models (e.g., ResNet, Waterbirds, CelebA, etc.), all of which are properly cited in the main text 5.2. License information and usage terms for these assets are respected.

    Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: The paper introduces a new benchmark dataset, AnimalsMetaCoCo, which is well documented. Details about its construction, structure, and intended use are provided in the main text 5.2 and supplementary material. Documentation will be made available alongside the dataset release.

Guidelines:

- The answer NA 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: [NA]

Justification: The paper does not involve any crowdsourcing or research with human subjects.

Guidelines:

- The answer NA 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: [NA]

Justification: The paper does not involve research with human subjects and therefore does not require IRB or equivalent approval.

Guidelines:

- The answer NA 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 rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core methodology and contributions of the paper do not involve the use of LLMs in any important, original, or non-standard way.

Guidelines:

- The answer NA 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 (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.