# Unified Multi-Layer Subspace Modeling for Cross-Domain OOD Detection

Gerhard Krumpl[1,2], Henning Avenhaus[2], and Horst Possegger[1]

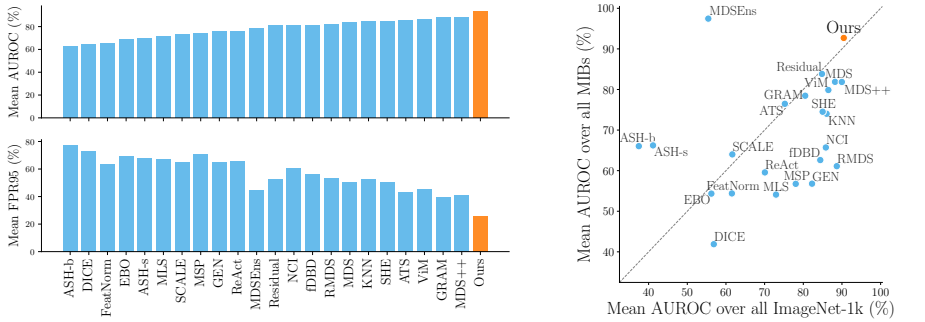
[1] Institute of Visual Computing, Graz University of Technology, Austria
gerhard.krumpl@icg.tugraz.at, possegger@tugraz.at
[2] KESTRELEYE GmbH, Austria

**Abstract.** Out-of-Distribution (OOD) detection remains a fundamental challenge for neural networks, whose predictions can be overconfident on inputs that deviate from the training distribution. Most post-hoc OOD detection methods derive scores from a single representation level (*e.g.*, logits or penultimate features) or combine multiple layers via depth selection or OOD-calibrated weighting. However, because OOD shifts are diverse, the most informative representation level can vary strongly across OOD types and domains, making fixed-layer choices and OOD-calibrated aggregation brittle. In this paper, we propose PRISM (Projected Representation with Intermediate-layer Subspace Modeling), a model-agnostic post-hoc OOD detection method that models a unified multi-layer feature representation rather than aggregating independently scored layers. PRISM fuses intermediate and deep features into a single hierarchical embedding, estimates an in-distribution (ID) principal subspace, and then combines two complementary signals: (i) a class-conditional Mahalanobis distance in the projected subspace and (ii) the residual energy orthogonal to the learned manifold. This simple design avoids OOD-tuned layer weighting while capturing both in-subspace semantic deviations and off-subspace anomalies. Across diverse benchmarks spanning natural images, medical imaging, and industrial visual inspection, PRISM achieves consistent state-of-the-art cross-domain OOD detection performance with a single default configuration across all evaluated domains and architectures. We further show that PRISM incurs minimal inference overhead, making it practical for real-world deployment.

**Keywords:** Out-of-Distribution Detection · Intermediate Layer Feature Fusion

## 1 Introduction

Machine learning models have demonstrated strong performance across diverse domains [29, 54, 60] and are now widely deployed in real-world systems. Despite their strong in-distribution (ID) accuracy, such models can produce highly confident yet incorrect predictions for inputs that deviate substantially from the training distribution, creating a mismatch between confidence and correctness that may lead to silent and potentially safety-critical failures [23, 49, 52]. This is




**(a)** Aggregate performance across all benchmarks (top: AUROC↑, bottom: FPR95↓).

**(b)** Cross-domain AUROC correlation across methods.

**Fig. 1:** Comprehensive evaluation of post-hoc OOD detection across natural, medical, and industrial benchmarks. (a) Mean AUROC (↑) and FPR95 (↓) averaged over all datasets, OOD splits, and architectures. (b) Cross-domain correlation between ImageNet-1k [14] (OpenOOD [79]) and OpenMIBOOD [19] performance (AUROC). Our PRISM (**orange**) consistently exhibits strong performance across domains.

particularly concerning in high-consequence applications, where erroneous predictions may directly affect safety or operational reliability [25, 27, 48]. For example, in medical imaging [19], models may fail to deliver accurate diagnoses due to scanner variability or unfamiliar instruments, while in industrial inspection [10, 33] systems may misclassify foreign materials as food products or fail to detect damage to machinery or infrastructure during quality control. These limitations motivate out-of-distribution (OOD) detection: identifying inputs outside the training distribution to improve deployment reliability.

Post-hoc OOD detection is particularly attractive in practice because it can be applied to pretrained models without retraining, thereby preserving ID accuracy and avoiding costly re-optimization. Accordingly, a wide range of post-hoc approaches have been proposed, leveraging predicted probabilities [26, 44], output logits [24, 43], or feature embeddings from the penultimate layer [38, 55, 63].

Our motivation arises from the observation that post-hoc OOD performance varies substantially across OOD type and application domains (see Fig. 1) [19, 33]. Recent benchmarks such as OpenMIBOOD [19] and ICONIC-444 [33] provide large-scale OOD evaluations in medical and industrial domains, respectively, and show that methods performing strongly on natural image benchmarks [79] often degrade markedly in these specialized settings.

A key driver of this variability is that post-hoc OOD detection methods typically rely on a single network layer for discrimination, yet the most informative layer depends strongly on the ID dataset and OOD type [19, 35]. Consequently, most existing methods largely neglect information in intermediate layers of deep neural networks, implicitly assuming that the final representation alone is sufficient for OOD detection. While deep layers emphasize representations highly specialized for the ID task, intermediate and shallow layers capture complemen-

tary information that can be critical for distinguishing diverse OOD characteristics, which may manifest differently as network depth increases. Other methods exist, but prior multi-layer methods either select a single layer [76], aggregate layer-wise scores [20,35,58], or calibrate aggregation/hyperparameters using synthetic or held-out OOD data [16, 38, 64], introducing additional complexity and OOD dependence. This raises a key question: how can intermediate representations be effectively integrated into a unified OOD detection framework that remains reliable across domains and OOD types without OOD-specific calibration or tuning?

To address this question, we propose **PRISM**, an architecture-agnostic post-hoc OOD detection method designed for cross-domain robustness. PRISM follows a simple fuse-then-model principle: instead of computing and aggregating layer-wise OOD scores, it first integrates intermediate representations from shallow to deep layers into a unified feature space and then performs statistical modeling once in that fused representation. The fused features are projected into a compact subspace learned solely from ID training data, and PRISM combines complementary distance- and residual-based signals into a single OOD score. This design avoids layer selection, layer-wise aggregation heuristics, and OOD-dependent calibration while retaining the robustness benefits of statistical feature modeling. Our main contributions are as follows:

- We show that layer-wise aggregation and OOD-based calibration is highly sensitive to the chosen validation OOD set, leading to unstable cross-domain behavior.
- We introduce a unified multi-layer subspace modeling principle that eliminates layer selection, score aggregation, and OOD-dependent calibration, while retaining the robustness benefits of statistical feature modeling.
- We evaluate PRISM across natural, medical, and industrial benchmarks under a shared configuration, achieving the best overall mean AUROC and the best average rank across datasets without OOD-based calibration.

## 2 Related Work

Out-of-distribution (OOD) detection methods are commonly categorized into training-based and post-hoc approaches [75, 79]. In this work, we focus on post-hoc OOD detection, which is particularly attractive in practice because it operates on pretrained models without retraining, thereby preserving ID accuracy and enabling scalable deployment [39, 79, 80].

Early post-hoc methods derive OOD scores directly from the model output. Probability-based approaches, such as the maximum softmax probability (MSP) [26], identify OOD samples by low prediction confidence. Subsequent works leverage logits directly, including the maximum logit score (MLS) [24] and energy-based OOD detection (EBO) [43]. Several model enhancement approaches further refine output-based detection by modifying activations or weights at inference time, for example, through activation clipping [61] or sparsification [15, 62].

Feature-based post-hoc methods instead model the distribution of deep representations and measure deviation using distance or density-based criteria. Mahalanobis distance-based detection [38] fits class-conditional Gaussian distributions to penultimate-layer features and uses the minimum distance to the class means as an OOD score. Several extensions refine this idea by normalizing or using relative scoring [55, 80], while nearest-neighbor approaches [63] exploit local structure in the embedding space. ViM [71] complements feature distances with residual information derived from a principal subspace. Kernel PCA [18] replaces linear PCA with non-linear kernel mappings in the penultimate feature space and uses reconstruction error or distance as an OOD score. Overall, these approaches use the model output, the penultimate layer, or both, thereby ignoring information from intermediate layers. PRISM instead explicitly leverages shallow-to-deep representations by modeling a fused feature hierarchy.

Existing multi-layer strategies typically leverage intermediate feature activations and compute layer-specific OOD scores, which are then aggregated. For example, ATS [35] models activation statistics across layers, while GRAM [58] derives OOD scores from deviations in layer-wise Gram matrices capturing second-order feature correlations. CORES [64] estimates consistency of representations across layers using statistical measures, whereas NAP [53] binarizes intermediate activations and computes Hamming distances to characterize deviations from ID patterns. Similarly, multi-layer Mahalanobis-based approaches [4] extend distance-based detection by estimating class-conditional statistics at multiple depths. While intermediate representations capture complementary information, these strategies introduce additional design choices, aggregation heuristics, or calibration procedures that often rely on OOD or synthetic data and are sensitive to domain shifts.

Recent large-scale OOD detection benchmarks across medical [19] and industrial [33] domains show that many post-hoc methods exhibit substantial performance variation outside the settings in which they were originally developed. In particular, a large fraction of OOD detectors are designed and validated on CIFAR [32] or ImageNet-based [14] benchmarks, where they often perform strongly, but show markedly reduced reliability when transferred to domain-specific scenarios. This suggests that many methods implicitly adapt to the statistical characteristics of the benchmarks on which they are optimized [34, 79]. In contrast, statistical and distance-based approaches that operate directly in feature space have been observed to be more stable across diverse settings, such as architectures and training methodologies, motivating methods that exploit feature geometry in a principled, domain-agnostic manner [34, 50].

Overall, prior post-hoc methods demonstrate the effectiveness of feature-space geometry and multi-layer representations, but they differ substantially in how they incorporate hierarchical information. In this work, we adopt a simple design choice that is underexplored in post-hoc OOD detection. Rather than selecting individual layers or aggregating per-layer scores, we jointly leverage representations from shallow to deep layers and model the dominant ID structure in a low-dimensional subspace. By deriving a single OOD score from this

unified representation, our approach avoids aggregation heuristics, OOD-based calibration, and per-benchmark tuning while retaining the robustness benefits of statistical feature modeling.

## 3 Method

We first formalize the OOD detection problem (Sec. 3.1) and review methods based on Mahalanobis distance as a widely used post-hoc baseline (Sec. 3.2). We then introduce PRISM (Sec. 3.3), which follows a *fuse-then-model* design: it constructs a unified multi-layer representation, models the in-distribution (ID) structure in a principal subspace, and derives a single OOD score from projected Mahalanobis distance and residual energy.

### 3.1 Problem Statement

We consider an image input space $\mathcal{X} \subset \mathbb{R}^{C\times H\times W}$ and a set of class labels $\mathcal{Y} = \{1, \dots, K\}$. Let $f : \mathcal{X} \to \mathbb{R}^K$ denote a classifier pretrained on an ID dataset $\mathcal{D}_{\text{ID}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where samples are drawn i.i.d. from the joint distribution $P_{\text{ID}}(X, Y)$ over $\mathcal{X}_{\text{ID}} \times \mathcal{Y}$. We refer to $\mathcal{X}_{\text{ID}} \subset \mathcal{X}$ as the in-distribution (ID) input space and denote its associated label space by $\mathcal{Y}_{\text{ID}} = \mathcal{Y}$. At deployment, the classifier may encounter inputs $\mathbf{x} \in \mathcal{X}$ that do not belong to $\mathcal{X}_{\text{ID}}$, defining the out-of-distribution (OOD) input space $\mathcal{X}_{\text{OOD}} = \mathcal{X} \setminus \mathcal{X}_{\text{ID}}$. We target the *semantic* OOD regime, where OOD samples belong to classes unseen during training, such that the corresponding label space $\mathcal{Y}_{\text{OOD}}$ satisfies $\mathcal{Y}_{\text{OOD}} \cap \mathcal{Y}_{\text{ID}} = \varnothing$.

OOD detection is formulated as a binary decision problem based on a scoring function $s : \mathcal{X} \to \mathbb{R}$. An OOD detector $H : \mathcal{X} \to \{\text{ID}, \text{OOD}\}$ is defined as:

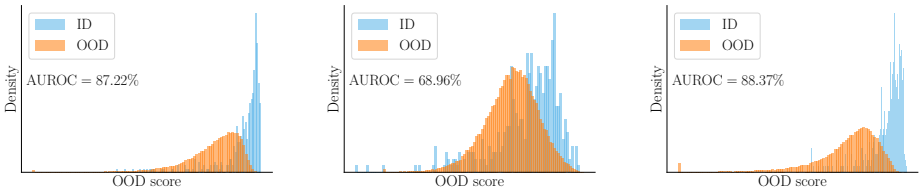
$$H(\mathbf{x}) = \begin{cases} \text{ID}, & \text{if } s(\mathbf{x}) \geq \tau \\ \text{OOD}, & \text{otherwise} \end{cases}, \tag{1}$$

where $\tau$ is a threshold typically chosen to retain a high fraction of ID samples (*e.g.*, 95% true positive rate).

### 3.2 Mahalanobis Distance for OOD Detection

We briefly review Mahalanobis-based OOD detection [38] to establish notation and motivate PRISM's reformulation in a unified fused feature space. Mahalanobis distance is a widely used post-hoc baseline that models ID features using class-conditional Gaussians. Let $\phi(\mathbf{x}) \in \mathbb{R}^M$ denote a feature representation extracted from a pretrained classifier, typically from the penultimate layer. Given the ID training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, the class-wise means $\widehat{\boldsymbol{\mu}}_c$ and shared covariance matrix $\widehat{\boldsymbol{\Sigma}}$ are estimated as

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



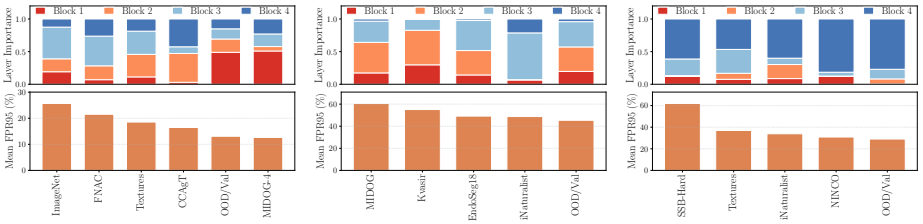

**(a)** MDS++ (Shallow) **(b)** MDS++ (Deep) **(c)** PRISM

**Fig. 2:** Comparison of OOD score distribution across layer depths. Histograms represent data sampled from PhaKIR (ID) and Cholec80 (Near-OOD) using a ResNet-18. (a) MDS++ applied to the first ResNet-18 block (shallow features). (b) MDS++ applied to the penultimate layer (deep features). (c) PRISM integrates the full architectural hierarchy (shallow to deep features).

where $N_c$ denotes the number of ID samples in class $c$ and $N$ the total number of ID samples. The Mahalanobis distance of a test sample $\mathbf{x}$ to class $c$ is

$$d_{\text{Maha}}(\mathbf{x}, c) = (\phi(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_c)^\top \, \widehat{\boldsymbol{\Sigma}}^{-1} \, (\phi(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_c)\,, \tag{3}$$

and the final OOD score is computed as the negative of the minimum distance to all class centroids. Although effective, Mahalanobis-based detection is sensitive to the choice of feature representation, as also shown in Fig. 2: applying it to shallow *vs*. deep layers yields markedly different ID-OOD separation across datasets. A common remedy is to compute Mahalanobis scores at multiple layers and aggregate them, either by averaging or by learning weights (*e.g.*, via logistic regression) using held-out or synthetic OOD data [4,38]. However, while such aggregation can improve performance on specific benchmarks, it is often sensitive to the chosen OOD validation distribution and may not generalize beyond it. As shown in Fig. 3, the layer weights learned by an OOD-calibrated aggregator (MDSEns [38]) vary substantially with the chosen OOD validation set: shallow layers are favored for some medical shifts (*e.g.*, MIDOG [7]), intermediate layers for others (*e.g.*, PhaKIR [56]), and deeper layers for ImageNet-1k [14]; even within the same domain, different near-OOD splits (*e.g.*, SSB-Hard [70] *vs*. NINCO [11]) lead to distinct optimal layer weights. Together, Figs. 2 and 3 indicate that the preferred representation depth varies substantially across domains or OOD definitions. This highlights a practical deployment trade-off: OOD-calibrated ensembles such as MDSEns can benefit from representative held-out OOD data for layer-weight calibration, but the resulting aggregation may be tied to the calibration distribution. To overcome the need for such calibration, we require an alternative that preserves information across depth without OOD-calibrated aggregation, which we realize in PRISM by fusing representations prior to statistical modeling. With PRISM, we thus prioritize transferable ID-only deployment by estimating all statistics from ID data and using a fixed configuration when representative OOD data are unavailable.




**(a)** MIDOG (OpenMIBOOD) **(b)** PhaKIR (OpenMIBOOD) **(c)** ImageNet-1k (OpenOOD)

**Fig. 3:** Sensitivity of layer aggregation to OOD validation data. For three benchmark tasks: (a) MIDOG, (b) PhaKIR, and (c) ImageNet-1k, we fit linear aggregation weights over ResNet blocks 1–4 using one OOD dataset as validation and evaluate the resulting model across all OOD test sets. The stacked bar plots (top) show the normalized layer weights learned for each calibration dataset, while the bars (bottom) indicate the corresponding mean FPR95 averaged over all OOD datasets.

### 3.3 PRISM: Projected Representation with Intermediate-layer Subspace Modeling

PRISM is based on a simple principle: instead of computing and aggregating layer-wise OOD scores, we first integrate intermediate representations across network depth into a single feature space and then apply statistical modeling once. The method consists of three steps: (i) construction of a unified multi-layer representation, (ii) approximating its ID structure using principal component analysis (PCA), and (iii) deriving an OOD score from projected Mahalanobis distance and residual energy. All feature statistics, PCA parameters, and class-conditional distributions are estimated using ID training data only. This formulation avoids selecting a single best layer, layer-wise aggregation heuristics, and OOD-dependent calibration.

**Multi-Layer Feature Construction.** Let $\{\phi^{(l)}(\mathbf{x})\}_{l=1}^{L}$ denote the feature maps extracted from $L$ selected intermediate layers of the classifier $f$. The structure of these activations depends on the underlying architecture. For convolutional neural networks (CNNs), $\phi^{(l)}(\mathbf{x}) \in \mathbb{R}^{C_l \times H_l \times W_l}$ represents spatial feature maps with $C_l$ channels, whereas for Vision Transformers (ViTs), $\phi^{(l)}(\mathbf{x}) \in \mathbb{R}^{T_l \times C_l}$ corresponds to a sequence of $T_l$ tokens of dimension $C_l$. To unify these representations while preserving global context, we apply an architecture-agnostic pooling operator $\mathcal{P}(\cdot)$ that aggregates over the non-channel dimensions:

$$\mathbf{z}^{(l)}(\mathbf{x}) = \mathcal{P}(\phi^{(l)}(\mathbf{x})). \tag{4}$$

In practice, we use global average pooling for CNN feature maps and token averaging for transformer layers, resulting in a single vector per layer. To prevent layers with larger activation magnitudes from dominating the joint representation, each pooled vector is $\ell_2$-normalized:

$$\widetilde{\mathbf{z}}^{(l)}(\mathbf{x}) = \frac{\mathbf{z}^{(l)}(\mathbf{x})}{\|\mathbf{z}^{(l)}(\mathbf{x})\|_2}. \tag{5}$$

Finally, the normalized representations are concatenated to form a single multi-layer feature vector:

$$\mathbf{z}(\mathbf{x}) = \left[\widetilde{\mathbf{z}}^{(1)}(\mathbf{x}), \ldots, \widetilde{\mathbf{z}}^{(L)}(\mathbf{x})\right] \in \mathbb{R}^D, \tag{6}$$

where $D = \sum_{l=1}^{L} C_l$ is the dimensionality of the joint feature space. This construction yields a single feature vector that jointly captures low-level structural and high-level semantic information. This per-layer normalization also reduces architecture- and dataset-dependent scale differences before PCA and subsequent scoring, improving the stability of the downstream projected-distance and residual-energy terms.

**Subspace Modeling of ID Data.** It is commonly assumed that neural network representations of ID data lie on a low-dimensional manifold and can be well-approximated by a low-dimensional subspace [71, 77]. In PRISM, we apply this assumption to our fused multi-layer representation $\mathbf{z}(\mathbf{x})$ and model its dominant ID variability with a linear subspace estimated by PCA. Given fused representations $\{\mathbf{z}(\mathbf{x}_i)\}_{i=1}^{N}$ from $\mathcal{D}_{\text{ID}}$, we compute the global mean $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{z}(\mathbf{x}_i)$ and derive the principal subspace $\mathbf{W} \in \mathbb{R}^{D\times d}$ from the top $d$ eigenvectors of the empirical covariance matrix. This subspace captures the dominant modes of variability of the ID data while discarding low-variance directions that are weakly supported by the training data.

At inference, a feature vector $\mathbf{z}(\mathbf{x})$ is decomposed into its projection $\mathbf{u}(\mathbf{x}) = \mathbf{W}^\top(\mathbf{z}(\mathbf{x}) - \boldsymbol{\mu})$ and its orthogonal residual:

$$\mathbf{r}(\mathbf{x}) = (\mathbf{z}(\mathbf{x}) - \boldsymbol{\mu}) - \mathbf{W}\mathbf{u}(\mathbf{x}). \tag{7}$$

Here, $\mathbf{u}(\mathbf{x})$ captures the alignment with the dominant ID semantic patterns, while $\mathbf{r}(\mathbf{x})$ captures feature components orthogonal to the ID manifold.

**Projected Mahalanobis Distance.** PRISM performs class-conditional Mahalanobis modeling directly in the unified projected space of the fused representation, yielding a single score without layer-wise scoring or aggregation. Using the projected coordinates $\mathbf{u}(\mathbf{x})$ of the ID training data, we estimate class-wise means $\widehat{\boldsymbol{\mu}}_c^u$ and a shared covariance matrix $\widehat{\boldsymbol{\Sigma}}^u$ in the PCA space. The projected Mahalanobis score for a sample $\mathbf{x}$ is then defined as

$$s_{\text{proj}}(\mathbf{x}) = \min_{c\in\mathcal{Y}} \left(\mathbf{u}(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_c^u\right)^\top \left(\widehat{\boldsymbol{\Sigma}}^u\right)^{-1} \left(\mathbf{u}(\mathbf{x}) - \widehat{\boldsymbol{\mu}}_c^u\right). \tag{8}$$

**Residual Energy.** Even if a sample is close to a class-conditional mean within the principal subspace, it may contain components that are not supported by the learned ID manifold. We capture this via the residual energy

$$s_{\text{res}}(\mathbf{x}) = \|\mathbf{r}(\mathbf{x})\|_2^2 = \|\mathbf{z}(\mathbf{x}) - \boldsymbol{\mu}\|_2^2 - \|\mathbf{u}(\mathbf{x})\|_2^2, \tag{9}$$

where the equality follows from $\mathbf{W}^\top\mathbf{W} = \mathbf{I}$. The residual energy measures how much of the representation cannot be explained by the ID subspace, serving as a complementary signal that captures novel or unsupported feature directions.

**Out-of-Distribution Detection with PRISM.** Both $s_{\text{proj}}(\mathbf{x})$ and $s_{\text{res}}(\mathbf{x})$ are nonnegative anomaly magnitudes. PRISM combines them into a single score:

$$s_{\text{PRISM}}(\mathbf{x}) = -\, s_{\text{proj}}(\mathbf{x}) \cdot s_{\text{res}}(\mathbf{x}) \tag{10}$$

This emphasizes samples that deviate from the ID structure both *within* and *outside* the principal subspace. We negate the product so that larger $s_{\text{PRISM}}(\mathbf{x})$ indicates more ID-like samples, consistent with Eq. (1). In contrast to weighted sums or learned ensembles, this combination introduces no balancing coefficients and requires no OOD-dependent calibration. We evaluate alternative score combinations in the supplementary material: additive fusion remains competitive but performs slightly worse than the multiplicative rule used in PRISM.

## 4 Experiments

We evaluate PRISM for cross-domain robustness and architectural generalization. Sec. 4.1 describes the evaluation protocol and benchmarks. In Sec. 4.2, we compare against 22 state-of-the-art post-hoc OOD detection methods across natural, medical, and industrial benchmarks [19,33,79]. Sec. 4.3 reports ablation results and analyzes the computational complexity.

### 4.1 Experimental Setup

**Benchmarks.** We evaluate PRISM on three large-scale OOD detection benchmarks: OpenOOD [74,78,79], OpenMIBOOD [19], and ICONIC-444 [33]. These benchmarks cover diverse acquisition settings and semantic granularity across natural images, industrial vision, histopathology, endoscopy, and MRI, enabling systematic evaluation of cross-domain robustness.

OpenOOD includes CIFAR-10/100 [32] and ImageNet-1k [14] as ID datasets. While CIFAR experiments are reported for completeness (see supplementary material), we focus on the large-scale ImageNet-1k task. OpenMIBOOD evaluates cross-domain generalization across three medical tasks: histopathology (MIDOG [7]), endoscopic video (PhaKIR [56]), and brain MRI (OASIS-3 [36]). ICONIC-444 evaluates OOD detection in high-resolution industrial vision systems and comprises four ID tasks (Almond, Wheat, Kernels, and Food-grade).

OpenOOD and OpenMIBOOD define near- and far-OOD splits based on benchmark-specific protocols, while ICONIC-444 also introduces extreme-OOD splits. To enable consistent cross-domain comparison, we standardize the evaluation into near-, far-, and extreme-OOD categories across all benchmarks, strictly adhering to official splits and extending them only where not explicitly defined. Across all tasks, we evaluate 21 OOD datasets spanning natural, medical, and industrial shifts (MNIST [37], FMNIST [72], Tiny ImageNet [66], SVHN [51], Textures [13], Places365 [81], SSB-Hard [70], NINCO [11], iNaturalist [69], OpenImage-O [71], CCAgT [3, 6], FNAC 2019 [57], Cholec80 [68], EndoSeg15 [12], EndoSeg18 [2], Kvasir [28], CATARACTS [1], ATLAS [40],

BraTS [8, 9, 47], MDS-Heart [5, 59, 65], CHAOS [30, 31]). A detailed overview of ID tasks, OOD datasets, and their grouping into near-, far-, and extreme-OOD categories is provided in the supplementary material.

**Architectures.** For the primary evaluation, we use the benchmark-defined backbones: ResNet-18 [22] for CIFAR-10/100, PhaKIR, and ICONIC-444; ResNet-50 is used for ImageNet-1k and MIDOG; and task-specific R(2+1)D [67] for the OASIS-3 brain MRI task. To evaluate architectural robustness, we additionally consider ViT-B/16 [17] and Swin-T [45] on ImageNet-1k, and CCT-7/7x2 [21] and ConvNeXt-S [46] on ICONIC-444.

**Layer Selection Strategy & Subspace Dimension.** We use an architecture-specific layer-selection rule: for each backbone, we extract features from standard stage/block outputs that are uniformly spaced across depth (stage outputs for convolutional networks; regularly spaced blocks for transformers; layers listed in the supplement). Unless stated otherwise, the PCA dimension is fixed to $d = 512$ for all experiments, independent of dataset, domain, or backbone. This avoids benchmark-specific tuning and keeps the protocol strictly ID-only across all evaluations.

**Evaluation Metrics.** Following established protocol [11,26,33,43,61,79], we report the false positive rate at recall 95% (FPR95) and the area under the receiver operating characteristic curve (AUROC). When reporting aggregated results, we (i) first average over the OOD test sets within each OOD category/task and, then (ii) average across categories/tasks, giving each category/task equal weight.

**Evaluation Protocol.** All inputs follow the benchmark's default preprocessing and are resized to the input shape required by the corresponding backbone. We compare PRISM against 22 state-of-the-art post-hoc OOD detection methods. For methods requiring ID statistics, we use the training split of the ID dataset. For methods requiring hyperparameter selection, we follow each benchmark's official validation protocol. MDSEns is the only method compared that requires OOD data for calibration; it is included for completeness (as a top-performing OpenMIBOOD [19] reference) and is marked accordingly in the results. Additional implementation details are provided in the supplementary material.

### 4.2 Benchmark Results

**Cross-Domain OOD Detection Performance.** Tab. 1 compares PRISM against 22 post-hoc OOD detection methods across all three domains. PRISM achieves the best overall mean AUROC (93.92%), improving over the second-best method (MDS++, 89.13%) by 4.79 points. It is also the only method that consistently remains among the top-performing approaches across all three domains. To further assess robustness beyond dataset averages, Fig. 4 reports AUROC averaged over near-, far-, and extreme-OOD categories. PRISM maintains high AUROC from near- to extreme-OOD, indicating robustness to increasing distributional shifts.

**Table 1:** AUROC (↑) performance across diverse imaging domains. Comparison of post-hoc OOD detection methods on natural (OpenOOD), medical (OpenMIBOOD), and industrial (ICONIC-444) benchmarks. Values represent the mean AUROC (%) across the benchmark-specific OOD test sets. Standard backbones are employed: ResNet-18 for CIFAR-10/100 (C10, C100), PhaKIR, and ICONIC-444; ResNet-50 for ImageNet-1k (IN-1k) and MIDOG; and R(2+1)D for OASIS-3. The **best** and second-best results in each column are highlighted in bold and underlined, respectively. * Results of MDSEns should be interpreted with caution, as discussed in Sec. 3.2 and in prior benchmark analyses [19].

| Method | OpenOOD | | | | OpenMIBOOD | | | | ICONIC-444 | | | | | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | C10 | C100 | IN-1k | **Mean** | MIDOG | PhaKIR | OASIS-3 | **Mean** | Almond | Wheat | Kernels | FG | **Mean** | |
| MSP [26] | 87.97 | 78.55 | 77.94 | 81.49 | 64.23 | 39.99 | 69.23 | 57.82 | 75.04 | 39.86 | 86.74 | 90.61 | 73.06 | 70.79 |
| MDS [38] | 88.66 | 63.83 | 89.83 | 80.77 | 78.14 | 73.66 | 98.73 | 83.51 | 97.00 | 98.64 | 96.65 | 88.96 | 95.31 | 86.53 |
| MDSEns* [38] | 91.79 | 78.24 | 54.41 | 74.82 | **96.87** | **98.49** | 99.89 | **98.41** | 93.38 | 98.44 | 95.82 | 87.59 | 93.81 | 89.01 |
| GRAM [58] | 91.78 | 85.00 | 85.61 | 87.46 | 89.49 | 36.46 | 99.42 | 75.12 | 97.77 | 98.70 | 97.82 | **95.89** | **97.55** | 86.71 |
| EBO [43] | 87.49 | 81.01 | 41.68 | 70.06 | 66.54 | 32.94 | 65.83 | 55.10 | 64.50 | 26.11 | 84.35 | 94.09 | 67.26 | 64.14 |
| RMDS [55] | 90.36 | 80.66 | 88.62 | 86.54 | 51.30 | 53.65 | 83.66 | 62.87 | 93.11 | 85.55 | 97.32 | 92.10 | 92.02 | 80.48 |
| ReAct [61] | 86.72 | 81.35 | 69.70 | 79.26 | 65.66 | 34.08 | 80.94 | 60.23 | 73.91 | 77.88 | 91.32 | 94.18 | 84.32 | 74.60 |
| MLS [24] | 87.43 | 80.95 | 70.21 | 79.53 | 65.71 | 32.94 | 65.87 | 54.84 | 65.08 | 26.11 | 84.35 | 93.65 | 67.30 | 67.22 |
| ViM [71] | 91.67 | 78.38 | 85.54 | 85.19 | 79.12 | 64.46 | 99.48 | 81.02 | 97.36 | 98.28 | 97.85 | 84.68 | 94.54 | 86.92 |
| KNN [63] | 91.49 | 80.87 | 85.64 | 86.00 | 76.17 | 51.12 | 99.28 | 75.52 | 94.16 | 88.82 | 97.03 | 89.87 | 92.47 | 84.66 |
| DICE [62] | 83.67 | 82.06 | 39.29 | 68.34 | 60.12 | 31.64 | 30.49 | 40.75 | 58.52 | 23.45 | 81.35 | 94.27 | 64.40 | 57.83 |
| FeatNorm [76] | 93.32 | 73.56 | 73.38 | 80.09 | 39.76 | 61.58 | 62.56 | 54.63 | 43.41 | 34.25 | 28.43 | 37.89 | 35.99 | 56.90 |
| ASH-b [15] | 72.99 | 81.08 | 26.17 | 60.08 | 68.94 | 51.10 | 84.37 | 68.14 | 66.86 | 33.33 | 74.22 | 91.12 | 66.38 | 64.87 |
| ASH-s [15] | 79.48 | 82.48 | 69.01 | 76.99 | 65.04 | 53.94 | 87.59 | 68.86 | 66.94 | 61.40 | 84.67 | 94.70 | 76.93 | 74.26 |
| Residual [71] | 86.41 | 52.10 | 84.23 | 74.24 | 80.93 | 76.44 | 98.84 | 85.40 | 96.24 | 98.72 | 95.32 | 54.68 | 86.24 | 81.96 |
| SHE [80] | 90.00 | 81.40 | 79.51 | 83.64 | 76.00 | 55.71 | 96.79 | 76.17 | 93.12 | 93.99 | 95.90 | 91.89 | 93.72 | 84.51 |
| GEN [44] | 88.84 | 81.17 | 87.33 | 85.78 | 63.38 | 41.01 | 69.23 | 57.87 | 75.18 | 40.69 | 87.64 | 93.85 | 74.34 | 72.66 |
| ATS [35] | 90.65 | 82.71 | 74.05 | 82.47 | 80.12 | 57.10 | 92.31 | 76.51 | 90.28 | 86.65 | 87.83 | 85.74 | 87.63 | 82.20 |
| SCALE [73] | 81.91 | 82.18 | 69.02 | 77.70 | 64.40 | 40.95 | 94.20 | 66.52 | 66.11 | 56.74 | 84.64 | 94.67 | 75.54 | 73.25 |
| fDBD [41] | 91.64 | 80.27 | 81.97 | 84.63 | 66.41 | 37.42 | 87.76 | 63.87 | 92.50 | 86.35 | 95.29 | 91.97 | 91.53 | 80.01 |
| MDS++ [50] | 91.08 | 80.53 | 89.88 | 87.16 | 80.72 | 71.69 | 99.88 | 84.09 | 97.04 | 98.79 | 98.12 | 90.54 | 96.12 | 89.13 |
| NCI [42] | 89.09 | 81.18 | 84.60 | 84.96 | 73.34 | 44.74 | 79.38 | 65.82 | 85.54 | 80.21 | 95.38 | 94.50 | 88.91 | 79.90 |
| **PRISM (Ours)** | **94.50** | **86.62** | **89.96** | **90.36** | 91.99 | 92.60 | **99.98** | 94.86 | **98.39** | **99.56** | **98.14** | 90.13 | 96.56 | **93.92** |

On OpenOOD, PRISM achieves the highest mean AUROC (90.36%), surpassing both GRAM and MDS++, which each perform strongly only on specific subsets (*e.g.*, CIFAR or ImageNet-1k). On OpenMIBOOD, PRISM attains 94.86%, substantially improving over MDS++ (84.09%), and emerges as the strongest ID-only alternative to the OOD-calibrated MDSEns (98.41%). On ICONIC-444, PRISM reaches 96.56%, closely matching the strongest industrial baseline (GRAM, 97.55%) and exceeding MDS++ (96.12%).

In contrast, competing approaches exhibit pronounced domain-dependent variability. For example, MDSEns achieves excellent results on OpenMIBOOD but drops to 54.41% on ImageNet-1k, consistent with the sensitivity of OOD-calibrated aggregation (*cf*. Fig. 3). GRAM performs strongly on ICONIC-444 (97.55%) yet degrades substantially on medical tasks (75.12%). Other methods, such as FeatNorm and DICE, show competitive performance on selected datasets but fail when transferred to other domains (*e.g.*, DICE: 30.49% on OASIS-3). Overall, these observations suggest that many post-hoc detectors are implicitly tailored to particular benchmark regimes and rely on design choices (*e.g.*, logit-only scoring, fixed-depth features, or OOD-calibrated aggregation) whose effectiveness varies across domains and OOD definitions. Interestingly, we further observe that several logit-based methods (*e.g.*, MLS, SCALE) degrade

**Table 2:** Ablation of PRISM components and PCA dimensionality. Performance is reported as AUROC (%) (↑) and FPR95 (%) (↓), averaged over all near-, far-, and extreme-OOD test sets across the three benchmarks. For OpenOOD, results are shown on the large-scale ImageNet-1k task. **(a)** Contribution of projected Mahalanobis distance ($s_{\text{proj}}$) and residual energy ($s_{\text{res}}$). **(b)** Sensitivity to the PCA projection dimension. **Best** and second-best results are highlighted in bold and underlined, respectively.

(a) Component ablation.

| $s_{\text{proj}}$ | $s_{\text{res}}$ | OpenOOD | | OpenMIBOOD | | ICONIC-444 | | Average | |
|---|---|---|---|---|---|---|---|---|---|
| | | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ |
| ✓ | | **89.94** | 38.62 | 91.42 | 32.32 | 95.21 | 16.05 | 92.19 | 29.00 |
| | ✓ | 76.39 | 58.20 | **95.72** | **17.21** | 93.82 | 20.30 | 88.64 | 31.90 |
| ✓ | ✓ | 89.88 | **35.83** | 94.86 | 21.10 | **95.67** | **14.02** | **93.47** | **23.65** |

(b) PCA dimensionality ablation.

| PCA-Dim | OpenOOD | | OpenMIBOOD | | ICONIC-444 | | Average | |
|---|---|---|---|---|---|---|---|---|
| | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ |
| 128 | 88.94 | 39.81 | 94.83 | 22.91 | 95.56 | 14.54 | 93.11 | 25.75 |
| 256 | 89.67 | 36.93 | **95.24** | 21.19 | 95.59 | 14.46 | **93.50** | 24.19 |
| 512 | 89.88 | **35.83** | 94.86 | **21.10** | 95.67 | 14.02 | 93.47 | **23.65** |
| 768 | **89.97** | 36.09 | 93.47 | 25.32 | **95.67** | **13.91** | 93.04 | 25.11 |

on extreme-OOD in some settings, underscoring that a larger semantic shift does not necessarily imply easier detection. PRISM instead uses a simple ID-only model of a fused shallow-to-deep representation, reducing opportunities for domain-specific optimization and yielding more consistent cross-domain behavior. This advantage persists beyond the mean AUROC: PRISM achieves the best overall average rank across datasets and shows significant improvements under paired per-dataset comparisons (see supplemental material, Fig. 1). To better understand the remaining errors, we provide a qualitative analysis of false positives in the supplement. Across OpenOOD, many failures arise from ID–OOD overlap, fine-grained semantic ambiguity, or context-dependent cases where the OOD object is visually close to, or only a small part of, an ImageNet-like scene. In medical and industrial benchmarks, errors often reflect subtle near-OOD confusions caused by acquisition artifacts or shared coarse shape, texture, and color attributes between ID and OOD categories. Overall, these failure modes are largely class-consistent and interpretable rather than arbitrary artifacts.

### 4.3 Ablation Study

**Contribution of PRISM components.** Tab. 2 analyzes the contribution of the projected Mahalanobis term ($s_{\text{proj}}$) and the residual energy term ($s_{\text{res}}$). The two components exhibit domain-dependent behavior: on OpenOOD and ICONIC-444, the projected term performs strongest, whereas on OpenMIBOOD, the residual term dominates, indicating that OOD deviations manifest differently across domains.

Even when used individually, either term already improves over classical penultimate-layer baselines when computed on the unified multi-layer representation. For comparison, averaged across all large-scale benchmarks, MDS obtains 87.00% AUROC (41.38% FPR95), MDS++ 88.37% (39.44%), and ViM 85.59% (41.79%), all below the single-component PRISM variants. This suggests that the gains do not arise merely from combining two scoring terms, but from modeling a fused representation that preserves complementary low-, mid-, and high-level cues. The projected subspace model then derives the OOD score from ID

**Table 3:** Performance comparison of post-hoc OOD detection methods across architectures. **(a)** OpenOOD (ImageNet-1k) using ResNet-50, Swin-T, and ViT-B/16. **(b)** ICONIC-444 using ResNet-18, CCT-7/7x2, and ConvNeXt-S. Methods are selected based on the mean AUROC across architectures within each benchmark.

(a) ImageNet-1k (OpenOOD)

| Method | ResNet-50 | | Swin-T | | ViT-B-16 | | Average | |
|---|---|---|---|---|---|---|---|---|
| | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ |
| MDS++ [50] | 89.88 | 42.34 | 88.85 | 52.15 | **89.52** | **40.37** | 89.42 | 44.96 |
| RMDS [55] | 88.62 | 47.71 | 88.02 | 55.59 | 88.38 | 53.87 | 88.34 | 52.39 |
| MDS [38] | 89.83 | 36.38 | 87.09 | 57.57 | 86.17 | 52.81 | 87.70 | 48.92 |
| ViM [71] | 85.54 | 41.42 | 88.15 | 49.50 | 83.45 | 50.45 | 85.71 | 47.12 |
| KNN [63] | 85.64 | 51.46 | 84.78 | 66.59 | 86.17 | 50.95 | 85.53 | 56.33 |
| NCI [42] | 84.60 | 52.40 | 86.15 | 61.21 | 84.62 | 66.59 | 85.13 | 60.07 |
| **PRISM** | **89.96** | **35.36** | **90.91** | **30.60** | 88.76 | 41.51 | **89.88** | **35.83** |

(b) ICONIC-444

| Method | ResNet-18 | | CCT-7/7x2 | | ConvNeXt-S | | Average | |
|---|---|---|---|---|---|---|---|---|
| | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ | AUROC ↑ | FPR95 ↓ |
| GRAM [58] | **97.55** | **8.41** | 88.75 | 34.64 | **96.29** | **15.93** | 94.19 | 19.66 |
| MDS++ [50] | 96.12 | 11.88 | 86.12 | 43.49 | 92.60 | 30.72 | 91.61 | 28.69 |
| ViM [71] | 94.54 | 22.11 | 82.56 | 49.84 | 92.99 | 28.44 | 90.03 | 33.46 |
| MDS [38] | 95.31 | 15.65 | 79.01 | 57.24 | 95.06 | 19.77 | 89.80 | 30.89 |
| SHE [80] | 93.72 | 23.65 | 82.29 | 49.38 | 88.27 | 39.21 | 88.09 | 37.41 |
| Residual [71] | 86.24 | 34.92 | 82.93 | 52.94 | 93.54 | 28.81 | 87.57 | 38.89 |
| **PRISM** | 96.56 | 9.87 | **94.65** | **16.08** | 95.79 | 16.12 | **95.67** | **14.02** |

feature geometry rather than a single layer, OOD-calibrated layer weights, or output confidence.

The two scores capture complementary geometric deviations. The projected Mahalanobis term detects violations of class-conditional structure within the principal ID subspace, while the residual term measures energy outside this subspace. Importantly, OOD deviations are not confined to residual directions: some samples lie outside the learned manifold, whereas others remain within it but violate class-conditional structure. In contrast to ViM, which models residual deviations at a single depth and combines them with a logit-based energy term for ID confidence, PRISM captures both in-subspace class-conditional deviations and off-subspace residual energy within a unified representation.

While each component can dominate on specific benchmarks, their combination achieves the highest overall average AUROC (93.47%) and the most balanced FPR95 (23.65%) across benchmarks, confirming that robust OOD detection requires modeling both types of deviation.

**Effect of Projection Dimension.** We further analyze the sensitivity to the PCA dimensionality. Performance remains stable across a broad range, with only minor variation in AUROC and FPR95 (see Tab. 2). While the selected value of $d = 512$ provides the best overall performance, the dependence on the projection dimension is generally weak, indicating that PRISM does not rely on precise tuning of subspace dimensionality. This further supports the claim that PRISM operates as a stable cross-domain method with weak sensitivity to $d$.

**Architectural Generalization.** To evaluate architectural robustness, we assess PRISM on diverse backbone families, including convolutional neural networks (ResNet-18/50 [22], ConvNeXt [46]) and transformer-based models (ViT-B/16 [17], Swin-T [45]). As shown in Tab. 3, PRISM consistently ranks among the top-performing methods across both OpenOOD and ICONIC-444, independent of the underlying architecture.

Although Tab. 3 reports only the strongest methods per benchmark for clarity (full results are provided in the supplementary material), notable rank variations remain across architectures and domains. The leading competitors are predominantly geometric and statistic-based approaches operating on penultimate

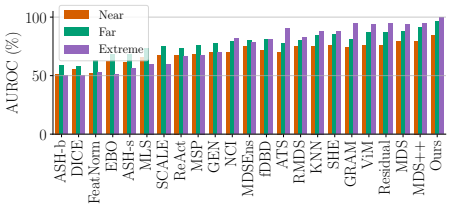



**Fig. 4:** AUROC (↑) averaged over near-, far-, and extreme-OOD categories across all three benchmarks. Methods are ordered by overall performance. All values are reported as percentages.

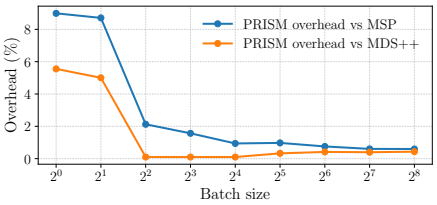



**Fig. 5:** Relative runtime overhead of PRISM on ImageNet-1k with ResNet-50. Runtime overhead as a percentage (%) of the total inference time compared to the baseline MSP and MDS++ methods.

features, which tend to generalize better across architectural families, consistent with prior findings [34,79]. Our results extend this observation by demonstrating stability not only across architectures but also across distinct domains.

Overall, PRISM maintains strong and consistent performance without specific tuning, confirming that unified multi-layer subspace modeling scales reliably across network designs.

**Computational Efficiency.** In PRISM, intermediate feature maps are spatially reduced via global pooling, transforming tensors of size $B \times C_l \times H_l \times W_l$ into $B \times C_l$. This compression scales with $\mathcal{O}(BH_lW_l)$ per layer but eliminates the dependence of the stacked representation $D = \sum_l C_l$ on the spatial resolution $(H, W)$. Consequently, all subsequent OOD scoring operations operate on resolution-independent feature vectors, ensuring that their computational cost does not increase with image size.

The additional computational cost of PRISM consists of three linear operations: (i) projection into a $d$-dimensional principal subspace with complexity $\mathcal{O}(BDd)$, (ii) Mahalanobis distance computation in the reduced space with $\mathcal{O}(BdK)$, and (iii) residual energy computation with $\mathcal{O}(BD)$, where $B$ denotes the batch size and $K$ the number of classes. Since $d \ll D$ and all computations are implemented as matrix multiplications, these operations scale linearly with $B$ and contribute only a bounded relative overhead compared to the backbone forward pass, which itself scales with $\mathcal{O}(BHW)$.

Overhead measures extra inference time beyond the backbone; MSP (max-softmax on logits) has near-zero overhead. For a fair comparison, we optimize the MDS++ implementation analogously to PRISM and evaluate all methods under identical measurement settings. Figure 5 reports the empirical runtime overhead on ImageNet-1k with ResNet-50. Compared to MSP and optimized MDS++, PRISM introduces less than 1.0% additional inference time for batch sizes $\geq 16$. Inference time is measured over 1k iterations, preceded by 200 warm-up runs on a synthetic input. These results indicate that multi-layer subspace modeling can be incorporated with negligible practical overhead while significantly boosting robustness.

## 5 Conclusion

We show that Out-of-Distribution (OOD) detection is influenced by representation depth: neither a single layer nor layer-wise score aggregation is consistently reliable across domains, and aggregation strategies calibrated on one OOD setting do not robustly transfer. This motivates preserving information across representation depths rather than relying on benchmark-specific depth preferences.

PRISM addresses this instability by fusing shallow-to-deep representations into a unified feature space and performing statistical modeling only after fusion, combining projected class-conditional Mahalanobis distance with orthogonal residual energy. As a result, PRISM avoids per-benchmark layer selection, aggregation heuristics, and OOD-based calibration, and delivers consistent state-of-the-art performance across natural, medical, and industrial benchmarks.

We use a fixed, architecture-aware layer sampling strategy and global pooling to ensure a consistent ID-only evaluation protocol. This design abstracts away spatial or token layouts, yielding a unified representation without architecture-specific fusion choices. More task-specific alternatives, such as tuned layer weights or spatially adaptive pooling, could improve individual benchmarks but would introduce additional priors and hyperparameters that may reduce transferability. Our minimal fusion design, therefore, isolates the central hypothesis of PRISM: unified multi-layer ID geometry can improve cross-domain OOD robustness without task-specific tuning or OOD-dependent calibration. Developing adaptive depth or pooling strategies that retain this transferability is an interesting direction for future work.

## Acknowledgements

We gratefully acknowledge support from KESTRELEYE GmbH, which provided the core infrastructure and resources throughout the project. HP was partially funded by the AI Ecosystems programme of the Austrian Federal Ministry of Innovation, Mobility and Infrastructure (BMIMI), managed by the Austrian Research Promotion Agency (FFG), under project DOMINO (923575).

## References

1. Al Hajj, H., Lamard, M., Conze, P.H., Roychowdhury, S., Hu, X., Maršalkaitė, G., Zisimopoulos, O., Dedmari, M.A., Zhao, F., Prellberg, J., Sahu, M., Galdran, A., Araújo, T., Vo, D.M., Panda, C., Dahiya, N., Kondo, S., Bian, Z., Vahdat, A., Bialopetravičius, J., Flouty, E., Qiu, C., Dill, S., Mukhopadhyay, A., Costa, P., Aresta, G., Ramamurthy, S., Lee, S.W., Campilho, A., Zachow, S., Xia, S., Conjeti, S., Stoyanov, D., Armaitis, J., Heng, P.A., Macready, W.G., Cochener, B., Quellec, G.: CATARACTS: Challenge on automatic tool annotation for cataRACT surgery. Med. Image Anal. (2019)
2. Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., Fuentes, F., Flouty, E., Mohammed, A., Pedersen, M., et al.: 2018 Robotic Scene Segmentation Challenge. arXiv preprint arXiv:2001.11190 (2020)

3. Amorim, J.G.A., Macarini, L.A.B., Matias, A.V., Cerentini, A., Onofre, F.B.D.M., Onofre, A.S.C., Von Wangenheim, A.: A novel approach on segmentation of agnor-stained cytology images using deep learning. In: Proc. CBMS (2020)
4. Anthony, H., Kamnitsas, K.: On the Use of Mahalanobis Distance for Out-of-distribution Detection with Neural Networks for Medical Imaging. In: Proc. UNSURE (2023)
5. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications (2022)
6. Atkinson Amorim, J.G., Matias, A., Bottamedi, T., Sanches, V., Costa, A.F., Onofre, F., Onofre, A., Wangenheim, A.: CCAgT: Images of Cervical Cells with AgNOR Stain Technique (2022)
7. Aubreville, M., Wilm, F., Stathonikos, N., Breininger, K., Donovan, T.A., Jabari, S., Veta, M., Ganz, J., Ammeling, J., van Diest, P.J., et al.: A comprehensive multi-domain dataset for mitotic figure detection. Scientific data (2023)
8. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. arXiv preprint arXiv:2107.02314 (2021)
9. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the Cancer Genome Atlas Glioma MRI Collections with Expert Segmentation Labels and Radiomic Features. Scientific data (2017)
10. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C.: The MVTec Anomaly Detection Dataset: A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection. In: IJCV (2021)
11. Bitterwolf, J., Mueller, M., Hein, M.: In or Out? Fixing ImageNet Out-of-Distribution Detection Evaluation. In: Proc. ICLR Workshops (2023)
12. Bodenstedt, S., Allan, M., Agustinos, A., Du, X., Garcia-Peraza-Herrera, L., Kenngott, H., Kurmann, T., Müller-Stich, B., Ourselin, S., Pakhomov, D., et al.: Comparative Evaluation of Instrument Segmentation and Tracking Methods in Minimally Invasive Surgery. arXiv preprint arXiv:1805.02475 (2018)
13. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing Textures in the Wild. In: Proc. CVPR (2014)
14. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proc. CVPR (2009)
15. Djurisic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely Simple Activation Shaping for Out-of-Distribution Detection. In: Proc. ICLR (2023)
16. Dong, X., Guo, J., Li, A., Ting, W.T., Liu, C., Kung, H.: Neural Mean Discrepancy for Efficient Out-of-Distribution Detection. In: Proc. CVPR (2022)
17. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: Proc. ICLR (2021)
18. Fang, K., Tao, Q., Lv, K., He, M., Huang, X., YANG, J.: Kernel PCA for Out-of-Distribution Detection. In: NeurIPS (2024)
19. Gutbrod, M., Rauber, D., Nunes, D.W., Palm, C.: OpenMIBOOD: Open Medical Imaging Benchmarks for Out-Of-Distribution Detection. In: Proc. CVPR (2025)
20. Haroush, M., Frostig, T., Heller, R., Soudry, D.: A Statistical Framework for Efficient Out of Distribution Detection in Deep Neural Networks. In: Proc. ICLR (2022)

21. Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., Shi, H.: Escaping the Big Data Paradigm with Compact Transformers. arXiv preprint arXiv:2104.05704 (2022)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: Proc. CVPR (2016)
23. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proc. CVPR (2019)
24. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling Out-of-Distribution Detection for Real-World Settings. In: Proc. ICML (2022)
25. Hendrycks, D., Carlini, N., Schulman, J., Steinhardt, J.: Unsolved Problems in ML Safety. arXiv preprint arXiv:2109.13916 (2021)
26. Hendrycks, D., Gimpel, K.: A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. In: Proc. ICLR (2017)
27. Hendrycks, D., Mazeika, M.: X-Risk Analysis for AI Research. arXiv preprint arXiv:2206.05862 (2022)
28. Jha, D., Smedsrud, P.H., Riegler, M.A., Halvorsen, P., De Lange, T., Johansen, D., Johansen, H.D.: Kvasir-seg: A segmented polyp dataset. In: MMM (2020)
29. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Hassabis, D.: Highly accurate protein structure prediction with alphafold. Nature (2021)
30. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., Baydar, B., Lachinov, D., Han, S., Pauli, J., Isensee, F., Perkonigg, M., Sathish, R., Rajan, R., Sheet, D., Dovletov, G., Speck, O., Nürnberger, A., Maier-Hein, K.H., Bozdağı Akar, G., Ünal, G., Dicle, O., Selver, M.A.: Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. Med. Image Anal. (2021)
31. Kavur, A.E., Selver, M.A., Dicle, O., Barış, M., Gezer, N.S.: CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data (2019)
32. Krizhevsky, A., Hinton, G.E.: Learning Multiple Layers of Features from Tiny Images. Master's thesis, Department of Computer Science, University of Toronto (2009)
33. Krumpl, G., Avenhaus, H., Possegger, H.: ICONIC-444: A 3.1-Million-Image Dataset for OOD Detection Research. In: Proc. WACV (2026)
34. Krumpl, G., Avenhaus, H., Possegger, H.: One Model, Many Behaviors: Training-Induced Effects on Out-of-Distribution Detection. In: Proc. WACV (2026)
35. Krumpl, G., Avenhaus, H., Possegger, H., Bischof, H.: ATS: Adaptive Temperature Scaling for Enhancing Out-of-Distribution Detection Methods. In: Proc. WACV (2024)
36. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., et al.: OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. medrxiv (2019)
37. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based Learning Applied to Document Recognition. IEEE **86**(11), 2278–2324 (1998)
38. Lee, K., Lee, K., Lee, H., Shin, J.: A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In: NeurIPS (2018)

39. Liang, S., Li, Y., Srikant, R.: Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks. In: Proc. ICLR (2018)
40. Liew, S.L., Lo, B.P., Donnelly, M.R., Zavaliangos-Petropulu, A., Jeong, J.N., Barisano, G., Hutton, A., Simon, J.P., Juliano, J.M., Suri, A., et al.: A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms. Scientific data (2022)
41. Liu, L., Qin, Y.: Fast decision boundary based out-of-distribution detector. In: ICML (2024)
42. Liu, L., Qin, Y.: Detecting Out-of-Distribution through the Lens of Neural Collapse. In: Proc. CVPR (2025)
43. Liu, W., Wang, X., Owens, J.D., Li, Y.: Energy-based Out-of-distribution Detection. In: NeurIPS (2020)
44. Liu, X., Lochman, Y., Christopher, Z.: Gen: Pushing the limits of softmax-based out-of-distribution detection. In: Proc. CVPR (2023)
45. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In: Proc. ICCV (2021)
46. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A ConvNet for the 2020s. In: Proc. CVPR (2022)
47. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J.S., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E.R., Weber, M., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, Ç., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M.T., Sarikaya, D., Schwartz, L.H., Shin, H., Shotton, J., Silva, C.A., Sousa, N.J., Subbanna, N.K., Székely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Ünal, G.B., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Leemput, K.V.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans. Med. Imaging (2015)
48. Mohseni, S., Wang, H., Yu, Z., Xiao, C., Wang, Z., Yadawa, J.: Practical Machine Learning Safety: A Survey and Primer. arXiv preprint arXiv:2106.04823 (2021)
49. Moosavi-Dezfooli, S.M., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: Proc. CVPR (2017)
50. Mueller, M., Hein, M.: Mahalanobis++: Improving OOD Detection via Feature Normalization. In: Proc. ICML (2025)
51. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading Digits in Natural Images with Unsupervised Feature Learning. NeurIPS Workshops (2011)
52. Nguyen, A.M., Yosinski, J., Clune, J.: Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In: Proc. CVPR (2015)
53. Olber, B., Radlak, K., Popowicz, A., M.Szczepankiewicz, Chachula, K.: Detection of out-of-distribution samples using binary neuron activation patterns. In: Proc. CVPR (2023)
54. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-Shot Text-to-Image Generation. In: Proc. ICML (2021)
55. Ren, J.J., Fort, S., Liu, J.Z., Roy, A.G., Padhy, S., Lakshminarayanan, B.: A Simple Fix to Mahalanobis Distance for Improving Near-OOD Detection. arXiv preprint arXiv:2106.09022 (2021)

56. Rueckert, T., Rauber, D., Maerkl, R., Klausmann, L., Yildiran, S.R., Gutbrod, M., Nunes, D.W., Moreno, A.F., Luengo, I., Stoyanov, D., Toussaint, N., Cho, E., Kim, H.B., Choo, O.S., Kim, K.Y., Kim, S.T., Arantes, G., Song, K., Zhu, J., Xiong, J., Lin, T., Kikuchi, S., Matsuzaki, H., Kouno, A., Manesco, J.R.R., Papa, J.P., Choi, T.M., Jeong, T.K., Park, J., Alabi, O., Wei, M., Vercauteren, T., Wu, R., Xu, M., Wang, A., Bai, L., Ren, H., Yamlahi, A., Hennighausen, J., Maier-Hein, L., Kondo, S., Kasai, S., Hirasawa, K., Yang, S., Wang, Y., Chen, H., Rodríguez, S., Aparicio, N., Manrique, L., Lyons, J.C., Hosie, O., Ayobi, N., Arbeláez, P., Li, Y., Al Khalil, Y., Nasirihaghighi, S., Speidel, S., Rueckert, D., Feussner, H., Wilhelm, D., Palm, C.: Comparative validation of surgical phase recognition, instrument keypoint estimation, and instrument instance segmentation in endoscopy: Results of the PhaKIR 2024 challenge. Med. Image Anal. (2026)
57. Saikia, A.R., Bora, K., Mahanta, L.B., Das, A.K.: Comparative assessment of CNN architectures for classification of breast FNAC images. Tissue and Cell (2019)
58. Sastry, C.S., Oore, S.: Detecting Out-of-Distribution Examples with Gram Matrices. In: Proc. ICML (2020)
59. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
60. Suhendar, H., Efelina, V., Ziveria, M.: Fruit Quality Classification using Convolutional Neural Network. JPCS (2022)
61. Sun, Y., Guo, C., Li, Y.: ReAct: Out-of-distribution Detection With Rectified Activations. In: NeurIPS (2021)
62. Sun, Y., Li, Y.: DICE: Leveraging Sparsification for Out-of-Distribution Detection. In: Proc. ECCV (2022)
63. Sun, Y., Ming, Y., Zhu, X., Li, Y.: Out-of-distribution Detection with Deep Nearest Neighbors. In: Proc. ICML (2022)
64. Tang, K., Hou, C., Peng, W., Chen, R., Zhu, P., Wang, W., Tian, Z.: CORES: Convolutional Response-based Score for Out-of-distribution Detection. In: Proc. CVPR (2024)
65. Tobon-Gomez, C., Geers, A.J., Peters, J., Weese, J., Pinto, K., Karim, R., Ammar, M., Daoudi, A., Margeta, J., Sandoval, Z.L., Stender, B., Zheng, Y., Zuluaga, M.A., Betancur, J., Ayache, N., Chikh, M.A., Dillenseger, J., Kelm, B.M., Mahmoudi, S., Ourselin, S., Schlaefer, A., Schaeffter, T., Razavi, R., Rhode, K.S.: Benchmark for Algorithms Segmenting the Left Atrium From 3D CT and MRI Datasets. IEEE Trans. Med. Imaging (2015)
66. Torralba, A., Fergus, R., Freeman, W.T.: 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. TPAMI (2008)
67. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A Closer Look at Spatiotemporal Convolutions for Action Recognition . In: Proc. CVPR (2018)
68. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Endonet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. IEEE Trans. Medical Imaging (2017)
69. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The INaturalist Species Classification and Detection Dataset. In: Proc. CVPR (2018)
70. Vaze, S., Han, K., Vedaldi, A., Zisserman, A.: Open-Set Recognition: a Good Closed-Set Classifier is All You Need? In: Proc. ICLR (2022)
71. Wang, H., Li, Z., Feng, L., Zhang, W.: ViM: Out-Of-Distribution with Virtual-logit Matching. In: Proc. CVPR (2022)

72. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv preprint arXiv:1708.07747 (2017)
73. Xu, K., Chen, R., Franchi, G., Yao, A.: Scaling for Training Time and Post-hoc Out-of-distribution Detection Enhancement. In: Proc. ICLR (2024)
74. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., PENG, W., Wang, H., Chen, G., Li, B., Sun, Y., Du, X., Zhou, K., Zhang, W., Hendrycks, D., Li, Y., Liu, Z.: OpenOOD: Benchmarking Generalized Out-of-Distribution Detection. In: NeurIPS Datasets and Benchmarks Track (2022)
75. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized Out-of-Distribution Detection: A Survey. arXiv preprint arXiv:2110.11334 (2021)
76. Yu, Y., Shin, S., Lee, S., Jun, C., Lee, K.: Block Selection Method for Using Feature Norm in Out-of-Distribution Detection. In: Proc. CVPR (2023)
77. Zaeemzadeh, A., Bisagno, N., Sambugaro, Z., Conci, N., Rahnavard, N., Shah, M.: Out-of-Distribution Detection Using Union of 1-Dimensional Subspaces. In: Proc. CVPR (2021)
78. Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Li, Y., Liu, Z., Chen, Y., Li, H.: Openood v1.5: Enhanced benchmark for out-of-distribution detection. arXiv preprint arXiv:2306.09301 (2023)
79. Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Li, Y., Liu, Z., Chen, Y., Li, H.: OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. J. of DMLR (2024)
80. Zhang, J., Fu, Q., Chen, X., Du, L., Li, Z., Wang, G., Liu, X., Han, S., Zhang, D.: Out-of-Distribution Detection based on In-Distribution Data Patterns Memorization with Modern Hopfield Energy. In: Proc. ICLR (2023)
81. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million Image Database for Scene Recognition. TPAMI (2017)