Optimal Zero-shot Regret Minimization for Selective Classification With Out-of-Distribution Detection

Eduardo Dadalto *[1]

Marco Romanelli[2]

[1]Helsing, Paris, France
[2]Computer Science Dept., Hofstra University, Hempstead, New York, USA

Abstract

Selective Classification with Out-of-Distribution Detection (SCOD) is a general framework that combines the detection of incorrectly classified in-distribution samples and out-of-distribution samples. Previous solutions for SCOD heavily rely on the choice of Selective Classification (SC) and Out-of-Distribution (OOD) detectors selected at test time. Notably, the performance of these detectors varies across different underlying data distributions. Hence, a poor choice can affect the efficacy of the SCOD framework. On the other hand, making an informed choice is impossible without samples from both in- and out-distribution. We propose an optimal zero-shot black-box method for SCOD that aggregates off-the-shelf detectors, is based on the principle of regret minimization, and provides an improvement on the worst-case performance. We demonstrate that our method achieves performance comparable to state-of-the-art methods in several benchmarks while also shielding the user from the burden of blindly selecting the SC and OOD detectors, optimally minimizing the regret and attaining reduced rejection risk.

1 INTRODUCTION

Classification with an abstention option has become a prominent strategy to make Deep Neural Network classifiers more trustworthy. In particular, the need to identify wrong predictions arising from in-distribution (in-d) and out-distribution (out-d) data has been the subject of extensive research in recent years, both in the fields of Selective Classification (SC) [Geifman and El-Yaniv, 2017, 2019, Granese et al., 2021] and Out-of-Distribution (OOD) detection [Liang et al., 2018, Sastry and Oore, 2020, Dadalto et al., 2022, Djurisic et al., 2023]. Selective Classification with Out-of-Distribution Detection (SCOD) [Xia and Bouganis, 2022, Narasimhan et al., 2024] has recently been proposed as a general framework for the detection of misclassified samples drawn from the training in-d $\mathbb{P}_{\text{in}}$, and samples coming from an out-d $\mathbb{P}_{\text{out}} \neq \mathbb{P}_{\text{in}}$. Narasimhan et al. [2024] introduces a *black-box* solution for SCOD when only samples from $\mathbb{P}_{\text{in}}$ are available. It provides a *plugin* framework that allows to combine off-the-shelf SC and OOD detection methods to achieve a Bayes-optimal rejector in the most constraining scenario.

Why do we need a principled solution based on regret minimization? *Previous work places the burden on the practitioner to decide which off-the-shelf scores to choose.* However, when the user does not have access to samples drawn from both $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{out}}$, no informed decision can be made. Indeed, the problem of binary detection has been shown to be very challenging, especially when a detector is required to perform well on multiple domains, with no side information about the underlying distributions [Lee and Barber, 2021, Fang et al., 2022, Pichler et al., 2024]. As highlighted in Li et al. [2023], worst-case risk minimization is essential for trustworthy systems, as it shields the user from the risk of picking detectors that may fail catastrophically on one or more given domains.

Thus, the **questions** we address in this paper are:

Q1 Can we make a more informed choice instead of blindly selecting a detector from off-the-shelf, achieving good SCOD performance while also reducing the worst-case rejection risk?

Q2 Can we accomplish the aforementioned goal with a zero-shot framework that allows us to reject samples in the wild without any additional training?

To illustrate our objective, consider Figure 1, which shows the worst-case, i.e. highest, Area Under the Risk-Coverage Curve (AUC-RC) attained by different OOD detectors when plugged in the framework of Narasimhan et al. [2024] and

*Work done while working at Université Paris-Saclay CNRS CentraleSupélec.

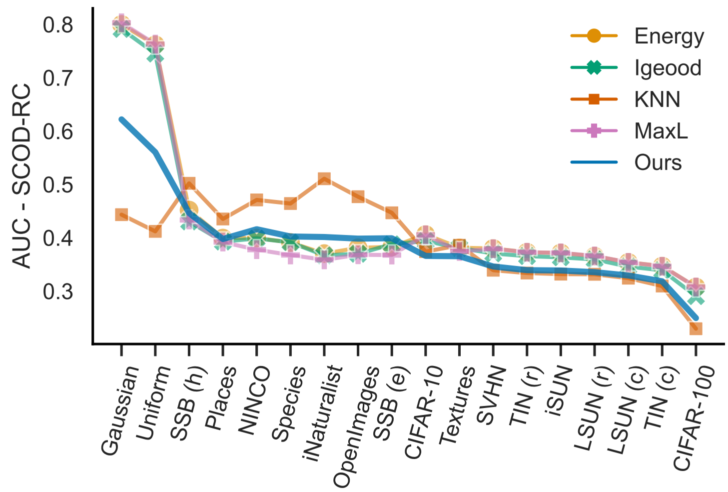


Figure 1: Worst-case performance for SCOD among 20 models, 3 in-d datasets (CIFAR-10, CIFAR-100, and ImageNet), and 18 out-d datasets. It highlights the absence of a singular superior performer in the task, exhibiting unpredictable efficacy depending on factors such as in- and out-domain scenarios and the OOD detection method applied within the SCOD framework. The proposed algorithm (blue line) strategically aggregates the off-the-shelf detectors in this example, do not require training, and mitigate catastrophic performance.

evaluated over 18 out-d domains. Clearly, not all the considered state-of-the-art (SOTA) detectors guarantee the same worst-case risk, e.g. Energy score (Energy) outperforms the others methods when the out-d domains are Gaussian or Uniform Noise [Hendrycks and Gimpel, 2017], but performs worse than the others when the out-d domains are, for instance, coming from curated datasets, such as Ninco [Bitterwolf et al., 2023], Species [Hendrycks et al., 2022], iNaturalist [Horn et al., 2017], or OpenImages [Krasin et al., 2017]. Had access to any out-d domain been granted, it would be possible to pick the most suitable method by, for instance, choosing the one that best performs on a given task. However, this is not possible in many realistic scenarios, where the only known domain is $\mathbb{P}_{\text{in}}$, and the only option is a black-box plugin estimator.

The principled zero-shot score we propose in this paper provides answers to the questions mentioned earlier. Our solution performs comparably to the state-of-the-art out-of-distribution detectors. Additionally, *it consistently reduces the worst-case AUC-RC across all tasks*, as indicated by the blue line in Figure 1 and in the main results in Section 5. In contrast, all other methods are impacted by the highest rejection risk in one or more cases.

This work makes the following **contributions**:

1. We identify and address a limitation in the existing SOTA black-box framework for SCOD. Without samples from the out-d, practitioners cannot make informed decisions on which off-the-shelf detection method to use (see Figure 1).
2. We provide a theoretical framework and derive a principled method to combine off-the-shelf OOD detectors in a zero-shot manner, meaning there is no need for any OOD data. Our approach is tailored to each input sample (see Section 3).
3. We compare the proposed method against a wide range of SOTA OOD detection methods on several benchmarks and models within the framework of Narasimhan et al. [2024]. The attained results are consistently comparable with the best-performing methods, while also reducing the worst-case rejection risk, which is crucial for the reliability of AI systems (see Section 5).

2 BACKGROUND

We consider the standard multi-class classification task, where, given a feature space $\mathcal{X} \in \mathbb{R}^d$, and a label space $\mathcal{Y} \doteq \{0, \ldots, C-1\} \subset \mathbb{N}$, a classifier is a function $f : \mathcal{X} \to \mathcal{Y}$ trained on a set of samples $S_n \doteq \{(\boldsymbol{x}_i, y_i)\}_{i=[n]}$ consisting of n i.i.d. training samples drawn according to the in-d $\mathbb{P}_{\text{in}}$ defined over the support $\mathcal{X} \times \mathcal{Y}$. In the usual setup, the training and test distribution coincide, i.e., $\mathbb{P}_{\text{te}} = \mathbb{P}_{\text{in}}$. Then, typically, for $\boldsymbol{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$, $f(\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} h_y(\boldsymbol{x}, \boldsymbol{\theta})$, where $h(\boldsymbol{x}, \boldsymbol{\theta}) : \mathcal{X} \to \mathbb{R}^C$, $h_y(\cdot, \boldsymbol{\theta})$ is the y-th component of $h(\cdot, \boldsymbol{\theta})$, and $\boldsymbol{\theta}$ is the vector of parameters that is fit to the training data by optimizing a loss function over S_n using an iterative algorithm such as Stochastic Gradient Descent (SGD).

Though this is an interesting theoretical setup, in practice, the ability to abstain from classifying a sample when the prediction confidence is expected to be low is a desideratum. This helps reduce the number of misclassifications for samples drawn from $\mathbb{P}_{\text{te}}$ that may be close to the learned decision boundary. Moreover, it helps dealing with situations in which $\mathbb{P}_{\text{te}} \neq \mathbb{P}_{\text{in}}$, e.g. $\mathbb{P}_{\text{te}} \doteq \pi_{\text{in}} \cdot \mathbb{P}_{\text{in}} + (1 - \pi_{\text{in}}) \cdot \mathbb{P}_{\text{out}}$ for a certain out-d $\mathbb{P}_{\text{out}}$, and a mixture parameter $\pi_{\text{in}} \in [0, 1]$.

SC is the problem of abstaining from classifying a sample, typically drawn from $\mathbb{P}_{\text{in}}$, when the classifier is not confident enough about its prediction Geifman and El-Yaniv [2017, 2019], Corbière et al. [2019], Liu et al. [2019], Huang et al. [2020], Granese et al. [2021]. The simplest way to model this goal is to consider a rejector $r : \mathcal{X} \to \{0, 1\}$, which is a binary function that outputs 1 when the prediction on $\boldsymbol{x} \in \mathcal{X}$ is rejected, and 0 otherwise. In this case, and for a given rejection budget $b_{\text{rej}} \in (0, 1)$, the optimal solution to the SC problem is $\min_{h,r} \mathbb{P}_{\text{in}}(y \neq h(\boldsymbol{x}), r(\boldsymbol{x}) = 0) : \mathbb{P}_{\text{in}}(r(\boldsymbol{x}) = 1) \leq b_{\text{rej}}$, where $\mathbb{P}_{\text{in}}(y \neq h(\boldsymbol{x}), r(\boldsymbol{x}) = 0)$ is the probability of accepting a prediction that is incorrect, and $\mathbb{P}_{\text{in}}(r(\boldsymbol{x}) = 1)$ is the probability of rejecting a prediction. Most of the SC frameworks in the literature consider the rejector to be a function of $h(\cdot, \boldsymbol{\theta})$, e.g. its output, and can be learned jointly with the classifier by adapting the training loss function Corbière et al. [2019], Huang et al. [2020]. In line with Narasimhan et al. [2024], we consider the solution in Geifman and El-Yaniv [2017], where the decision is made by comparing the Max Soft Probability (MSP) of a pre-trained model $h(\cdot, \boldsymbol{\theta})$ to a threshold.

OOD detection is the problem of detecting samples drawn from a distribution $\mathbb{P}_{\text{out}}$ that is different from the training distribution $\mathbb{P}_{\text{in}}$, for instance when $\mathbb{P}_{\text{te}} \doteq \pi_{\text{in}} \cdot \mathbb{P}_{\text{in}} + (1 - \pi_{\text{in}}) \cdot \mathbb{P}_{\text{out}}$. In this case, the optimal solution to the OOD problem is given by $\min_r \mathbb{P}_{\text{te}}(r(\boldsymbol{x}) = 0) : \mathbb{P}_{\text{in}}(r(\boldsymbol{x}) = 1) \leq b_{\text{fpr}}$, where $b_{\text{fpr}} \in (0, 1)$ is the False Positive Rate (FPR) budget, i.e., the fraction of in-d samples incorrectly predicted as out-d. In line with the plugin framework, we consider the most popular SOTA solutions for this problem, i.e. methods that consider the rejector to be a function of the pre-trained classifier, such as Liang et al. [2018], Liu et al. [2020], Feng et al. [2022], among others.

SCOD is the framework that combines SC and OOD. According to Narasimhan et al. [2024], the optimal solution to the SCOD problem is given by

$$\min_{h,r} \Big[(1 - c_{\text{fn}}) \cdot \mathbb{P}_{\text{in}}(y \neq h(\boldsymbol{x}), r(\boldsymbol{x}) = 0) + c_{\text{fn}} \cdot \mathbb{P}_{\text{out}}(r(\boldsymbol{x}) = 0) : \mathbb{P}_{\text{te}}(r(\boldsymbol{x}) = 1) \leq b_{\text{rej}} \Big], \quad (1)$$

where $c_{\text{fn}} \in [0, 1]$ is a user-specified cost of not rejecting an out-d sample.

A way to deal with it would be to perform hypothesis testing on the true distributions by defining s^*_{sc} and s^*_{ood} where

$$s^*_{\text{sc}}(\boldsymbol{x}) \doteq \max_{y \in [C]} \mathbb{P}_{\text{in}}(y \mid \boldsymbol{x}), \ s^*_{\text{ood}}(\boldsymbol{x}) \doteq \frac{\mathbb{P}_{\text{in}}(\boldsymbol{x})}{\mathbb{P}_{\text{out}}(\boldsymbol{x})}, \quad (2)$$

and comparing them with a threshold. Clearly, the two quantities above require full knowledge of $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{out}}$. The function h and r could be optimized for Equation (1) if we had samples from $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{out}}$, or a way to estimate the underlying mixture $\mathbb{P}_{\text{te}}$. However, in this work, and in line with the black-box solution presented in Narasimhan et al. [2024] (cf. Equation (3)), we do not assume access to out-d samples from $\mathbb{P}_{\text{out}}$, and we consider the classifier to be a pre-trained one, and we seek to leverage existing selective classification and OOD detection techniques to estimate the quantities in Equation (2).

Plugin Estimator. Given $s_{\text{sc}}(\cdot)$ and $s_{\text{ood}}(\cdot)$ scores as an estimate of $s^*_{\text{sc}}(\cdot)$, $s^*_{\text{ood}}(\cdot)$ respectively, derived from one SC method, and one OOD method among those listed in Section 7, the black-box plugin estimator takes the following form:

$$r_{\text{BB}}(\boldsymbol{x}) = \mathbb{1}[((1 - c_{\text{in}} - c_{\text{out}}) \cdot s_{\text{sc}}(\boldsymbol{x}) + c_{\text{out}} \cdot \beta(s_{\text{ood}}(\boldsymbol{x})) < t_{\text{BB}})], \quad (3)$$

where $c_{\text{in}}, \ c_{\text{out}} \in [0, 1]$, $\beta(s_{\text{ood}}(\cdot)) = -1/s_{\text{ood}}(\cdot)$, $t_{\text{BB}} = 1 - 2 \cdot c_{\text{in}} - c_{\text{out}}$, and Equation (3) coincides with Narasimhan et al. [2024, Lemma 3.1].

3 ZERO-SHOT SCOD WITH MINIMIZED REGRET

The SOTA black-box plugin method in Narasimhan et al. [2024] presents an optimal way to combine SC and OOD detection, using off-the-shelf methods to obtain the needed scores.

Crucially, a limiting aspect is that it does not provide a way to pick which detectors to use. In this work, we propose a way to aggregate multiple SOTA OOD detectors, according to the information theoretical notion of *regret minimization* Barron et al. [1998]. Notably, we propose a framework to combine multiple OOD detection scores into one, in a zero-shot way, without training a new detector.

As we shall see,

- Our solution achieves performance comparable to the best OOD detector in each considered task;
- Most importantly, it allows the user to aggregate multiple detectors from the literature, rather than choose one, which may have varying performance when evaluated on different domains.

3.1 PROBABILISTIC DETECTION SCORE (PDS): FROM A DISTANCE-BASED TO A CONFIDENCE-BASED OOD DETECTION

As defined in Equation (2), OOD detection scores can be summarized as a scalar function that tries to estimate the quantity $s_{\text{ood}}(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}^+$ such that $s_{\text{ood}}(\boldsymbol{x}) \approx (\mathbb{P}_{\text{in}}/\mathbb{P}_{\text{out}})(\boldsymbol{x})$, also referred to as likelihood ratio.

Low values of s indicate that the example is likely to be sampled from an out-d, while high values indicate otherwise. In the OOD literature, this formulation is also referred to as *confidence-based* scores. On the other hand, a popular interpretation of the OOD detection problem follows a *distance-based* setting, where higher values of the scores would indicate that the sample is far from the in-d, making it incompatible with confidence-based detection frameworks.

To allow any score function to be correctly plugged in our framework, that we shall present in Section 3, we propose a Probabilistic Detection Score (PDS) transformation to ground the scores into a common range and same distribution for in-d samples. This transformation is based on the empirical estimate of the Cumulative Distribution Function (CDF) which converges almost surely to the true CDF. As a result, we define PDS as being the function

$$\bar{s}_{\text{ood}} : \mathcal{X} \to [0, 1], \tag{4}$$

which unifies confidence-based and distance-based scores without loss of detection performance.

Algorithm 1 Probabilistic Detection Score (PDS) Transformation

Input: sorted set (ascending order) of confidence-based soft scores of m in-d samples $S_m = \{\delta : s_{\text{ood}}(\boldsymbol{x}_1) = \delta_1 \leq \cdots \leq \delta_m = s_{\text{ood}}(\boldsymbol{x}_m)\}$ and the score to be transformed $s_{\text{ood}}(\boldsymbol{x})$.

// new lower and upper bounds, respectively
$\delta_0 \leftarrow \delta_1/m, \quad \delta_{m+1} \leftarrow \delta_m \cdot m$

// concatenate bounds to original scores vector
$S \leftarrow \delta_0 \oplus S_m \oplus \delta_{m+1}$
$C \leftarrow [1/(m+2) \;\; 2/(m+2) \;\; \ldots \;\; (m+2)/(m+2)]$

// lower and upper query indexes
$i \leftarrow \arg\min_{k \in [0 \ldots m]} |s_{\text{ood}}(\boldsymbol{x}) - S_k|, \quad j \leftarrow i + 1$

// interpolation
$\bar{s}_{\text{ood}}(\boldsymbol{x}) \leftarrow C_i + (s_{\text{ood}}(\boldsymbol{x}) - S_i) \frac{C_j - C_i}{S_j - S_i}$

Return: $\bar{s}_{\text{ood}}(\boldsymbol{x})$

Algorithm 1 is an implementation of a quantile transformation that shows how to obtain the PDS for a confidence-based score. For a distance-based score, it changes slightly: first, the scores are flipped around the zero-value (i.e., multiplied by -1) to align with confidence-based scores; and finally, the heuristics for the lower and upper bounds changes: the division and multiplication operations in Algorithm 1 are swapped as we are dealing with strict negative values this time (i.e., $\delta_0 = \delta_1 \cdot m$ and $\delta_{m+1} = \delta_m/m$). After this pre-processing, Algorithm 1 can be applied as is.

The output of the PDS algorithm, transforms the s_{ood} scores into the corresponding value on the CDF curve. Intuitively, the empirical CDF, estimated through the sorted scores of samples from $\mathbb{P}_{\text{in}}$, maps a sample to its probability of being in-d. Indeed, considering Figure 2b, a sample will have a matching high $\bar{s}_{\text{ood}}$ score if its s_{ood} score is on the in-d side of Figure 2a, i.e. if it is likely in-d, and far from the error area where the histograms of in-d and out-d overlap.

Using this interpretation, the $\bar{s}_{\text{ood}}$ score can be regarded as a measure of the likelihood of being in-d, and used to define a detection method based on a distribution $\mathbb{Q}_{Z|\boldsymbol{x}}$ where Z is a binary random variable indicating whether the sample $\boldsymbol{x}$ is likely to be in-d if it takes the value 0 or out-d if it takes the value 1. Usually, a *Hard Rejector* is derived by comparing s to a real-valued hyperparameter γ, i.e., $r'(\boldsymbol{x}) = \mathbb{1}[s(\boldsymbol{x}) \leq \gamma]$. In our case, the PDS can be compared to a confidence value $\alpha \in [0, 1]$ that will indicate the desired False Negative Rate (FNR), or

$$r(\boldsymbol{x}) = \mathbb{1}[\bar{s}_{\text{ood}}(\boldsymbol{x}) \leq \alpha]. \tag{5}$$

Figure 2 showcases the PDS transformation for a confidence-based score (energy) and a distance-based score (KNN). We can observe the histograms of s_{ood} in Figure 2a and $\bar{s}_{\text{ood}}$ in Figure 2b, which maps any in-distribution to a uniform distribution in the unit range. OOD samples are mapped to low values due to the PDS transformation. The detection capacity is untouched as observed in Figure 2c, where the ROC curves of the PDS do not deviate from the ROC curve of the raw soft-score.

We plot the quantization error with 90% confidence bounds, showing that, with more than 10 in-distribution samples the absolute error between the PDS ROC and the original ROC is virtually zero. Furthermore, the study presented in Table 1 reports the average AURC for the proposed method across the OOD datasets considered in Section 4, using different values of m in Algorithm 1, specifically $\{5, 10, 20, 1000\}$: the values fluctuate slightly, the variance across the four choices of m is small, supporting the analysis of the m parameter shown in Figure 2.

Thus, through PDS, we can represent any score within any range into a score that resembles a distribution, requires few data points , is inexpensive to compute, and will allow the minimum regret principle introduced in Section 3.2 to aggregate multiple scores for the first time in OOD detection, enabling a more reliable SCOD setup. As a matter of fact, the detector provider can take care of this transformation, so that no data is needed to deploy our framework.

Model (in-d dataset)	Value of the parameter m				Variance
	5	10	20	1000	
VGG-16 (Cifar-10)	0.219	0.227	0.224	0.227	1.425E-5
DenseNet-121 (Cifar-100)	0.321	0.320	0.322	0.326	6.917E-6
ResNet-101 (ImageNet)	0.305	0.301	0.301	0.305	5.333E-6

Table 1: Ablation study of the parameter m in terms of average AURC.

3.2 Minimum Regret Probabilistic Score Aggregation (MRPSA)

Let us consider a model h as defined above and a score $\bar{s}_{\text{ood}}$ as described in Section 3.1, representing a probabilistic score. Usually, in practice, given a sample $\boldsymbol{x}$, a score is not only a function of the sample but also the underlying classifier model, i.e., we have $\bar{s}_{\text{ood}}(\boldsymbol{x}, h)$. Hence, the corresponding probability distribution $\mathbb{Q}_{Z|\boldsymbol{x},h}$ also depends on it. For the sake of simplicity, and assuming that a given pre-trained classifier is fixed, we will omit the dependence on the model in the following, using, without loss of generality, the notation $\bar{s}_{\text{ood}}(\boldsymbol{x})$ and $\mathbb{Q}_{Z|\boldsymbol{x}}$.

Now, consider a set of K soft-rejectors as in Section 3.1 and the set of corresponding distributions $\mathcal{Q} \doteq \{\mathbb{Q}^k_{Z|X}\}_{k=1}^K$, and let us assume that each rejector r_k is effective against at least one benchmark, i.e., it can successfully detect samples from $\mathbb{P}_{\text{in}}$ and $\mathbb{P}^k_{\text{out}}$ with high confidence at least for a given setting of $\mathbb{P}_{\text{in}}$ and $\mathbb{P}^k_{\text{out}}$. Notice that this is a mild assumption since all the published literature showcases a set of benchmarks where the proposed detectors are SOTA or close to it. Fixed an input sample $\boldsymbol{x}$, we would like to formally define $\mathbb{Q}^\star_{Z|\boldsymbol{x}}$ that performs well simultaneously over all the possible $|K|$ detection problem settings. This can be framed as the following problem:

$$\mathcal{L}(\mathcal{Q}, \boldsymbol{x}) = \min_{\mathbb{Q}_{Z|\boldsymbol{x}}} \max_{k\in\mathcal{K}} \mathbb{E}_{\mathbb{Q}^k_{Z|\boldsymbol{x}}}\left[-\log \mathbb{Q}_{Z|\boldsymbol{x}}\right], \tag{6}$$

which requires solving (6) for $\mathcal{Q}$ and for each given input sample $\boldsymbol{x}$. It is important to note that the minimization is performed over all distributions $\mathbb{Q}_{Z|\boldsymbol{x}}$, including elements that are not part of the set $\mathcal{Q}$. As it turns out, Equation (6) is computationally intractable. In Appendix A.1.1, we show how to obtain the following upper bound for the maximization problem in Equation (6). For any arbitrary choice of $\mathbb{Q}_{Z|\boldsymbol{x}}$, it holds that:

$$\max_{k\in\mathcal{K}} \mathbb{E}_{\mathbb{Q}^k_{Z|\boldsymbol{x}}}\left[-\log \mathbb{Q}_{Z|\boldsymbol{x}}\right] \leq \underbrace{\max_{k\in\mathcal{K}} \mathbb{E}_{\mathbb{Q}^k_{Z|\boldsymbol{x}}}\left[-\log \mathbb{Q}^k_{Z|\boldsymbol{x}}\right]}_{=\text{constant w.r.t. } \mathbb{Q}_{Z|\boldsymbol{x}}} + \underbrace{\max_{k\in\mathcal{K}} \mathbb{E}_{\mathbb{Q}^k_{Z|\boldsymbol{x}}}\left[\log\left(\frac{\mathbb{Q}^k_{Z|\boldsymbol{x}}}{\mathbb{Q}_{Z|\boldsymbol{x}}}\right)\right]}_{=\text{average worst-case regret Barron et al. [1998]}}. \tag{7}$$

According to the proof in Appendix A.1.2, this upper bound allows us to optimize the objective in Equation (6) by defining the surrogate objective in Equation (8).

$$\tilde{\mathcal{L}}(\mathcal{Q}, \boldsymbol{x}) = \min_{\mathbb{Q}_{Z|\boldsymbol{x}}} \max_{k\in\mathcal{K}} \mathbb{E}_{\mathbb{Q}^k_{Z|\boldsymbol{x}}}\left[\log\left(\frac{\mathbb{Q}^k_{Z|\boldsymbol{x}}}{\mathbb{Q}_{Z|\boldsymbol{x}}}\right)\right] = \min_{\mathbb{Q}_{Z|\boldsymbol{x}}} \max_{P_\Omega} \mathbb{E}_\Omega\left[D_{\text{KL}}\left(\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}} \| \mathbb{Q}_{Z|\boldsymbol{x}}\right)\right], \tag{8}$$

where the $\min$ is taken over all the possible distributions $\mathbb{Q}_{Z|\boldsymbol{x}}$; and Ω is a discrete random variable with P_Ω denoting a generic probability distribution whose probabilities are $(\omega_1, \ldots, \omega_{|\mathcal{K}|})$, i.e., $P_\Omega(k) = \omega_k$; and $D_{\text{KL}}(\cdot\|\cdot)$ is the Kullback–Leibler divergence (KL divergence), representing the expected value of regret of $\mathbb{Q}_{Z|U}$ w.r.t. the worst-case distribution in $\mathcal{Q}$. Finally, according to the proof in Appendix A.1.3, and utilizing the convex nature of the KL divergence, the solution to Equation (8) provides the optimal distribution $P^\star_\Omega$, i.e. the collection of weights $\{\omega^\star_k\}$, which leads to our soft-detector [Barron et al., 1998, Granese et al., 2024] $\mathbb{Q}^\star_{Z|\boldsymbol{x}}$:

$$\mathbb{Q}^\star_{Z|\boldsymbol{x}} = \sum_{k\in\mathcal{K}} \omega^\star_k \cdot \mathbb{Q}^k_{Z|\boldsymbol{x}}, \; P^\star_\Omega = \arg\max_{\{\omega_k\}} I_{\boldsymbol{x}}(\Omega; Z). \tag{9}$$

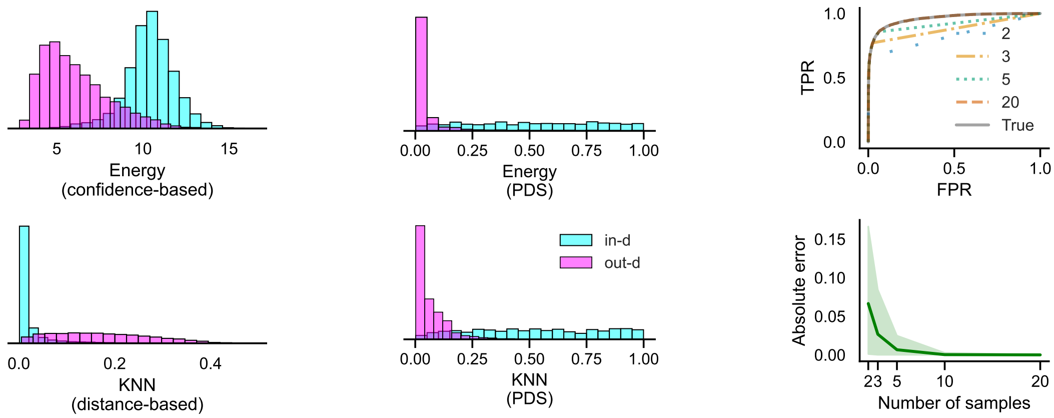
In Equation (9), $I_{\boldsymbol{x}}(\cdot;\cdot)$ denotes the Shannon mutual information between the random variable Ω, distributed according to $\{\omega_k\}$, and the binary soft-prediction variable Z, distributed according to $\mathbb{Q}^k_{Z|\boldsymbol{x}}$ and conditioned on the particular test example $\boldsymbol{x}$. The optimal combination of weights can be estimated by means of the Blahut-Arimoto Arimoto [1972] iterative algorithm that maximizes the mutual information in (9), parametrized by the weights $\{\omega_k\}$, with respect to the weights themselves. In line with Section 3.1, we can extract the score $\bar{s}_{\text{ood}}(\boldsymbol{x})$ from $\mathbb{Q}^\star_{Z|\boldsymbol{x}}$, i.e.,

$$r(\boldsymbol{x}) = \mathbb{1}\left[\mathbb{Q}^\star_{Z|X}(0|\boldsymbol{x}) < \alpha\right], \tag{10}$$

where $Z = 0$ indicates the event of detecting $\boldsymbol{x}$ as an in-d sample and use it in the black-box SCOD plugin framework of Narasimhan et al. [2024].

4 EXPERIMENTAL SETUP

Baselines. We consider the following post-hoc detection methods as off-the-shelf baselines: MSP [Hendrycks and



(a) Soft detection scores. (b) PDS scores. (c) ROC approximation.

Figure 2: MRPSA allows to move from confidence-based and distance-based soft detection scores to an effective probabilistic detection score with good convergence rate, using a relatively low value for the parameter m in Algorithm 1.

Gimpel, 2017], Energy [Liu et al., 2020], Mahalanobis (Maha) [Lee et al., 2018], Igeood [Dadalto et al., 2022], MaxCosine (MCos) [Techapanurak et al., 2020], ReAct [Sun et al., 2021], ODIN [Liang et al., 2018], Maximum logits (MaxL) [Hendrycks et al., 2022], KL-divergence matching (KL-M) [Hendrycks et al., 2022], Doctor [Granese et al., 2021], Relative Mahalanobis distance (RMaha) [Fort et al., 2021], and KNN [Sun et al., 2022]. Following popular OOD detection settings Fort et al. [2021], we followed the hyperparameter selection procedure suggested in the original papers, we used only the penultimate layer or logits outputs, and they are implemented so that they only have access to in-d data. In so doing, we extend the analysis reported in Figure 1 by considering a larger number of OOD detectors for SCOD, that we aggregate through MRPSA.

Models. We consider both pre-trained models and models trained from scratch with the following architectures: Residual Convolutional Neural Networks (ResNet) [He et al., 2016], Vision Transformers (ViT) [Dosovitskiy et al., 2021], MobileNet [Howard et al., 2017], DenseNet [Huang et al., 2017], and VGG [Simonyan and Zisserman, 2015]. We *do not* include any OOD data during training.

Datasets. The CIFAR-10 (C-10) dataset [Krizhevsky et al., 2009] comprises 32x32 pixel natural images categorized into 10 distinct classes, such as airplanes, ships, birds, and more. Similarly, the CIFAR-100 (C-100) dataset consists of natural images akin to those in CIFAR-10 but spanning 100 categories non-overlapping with C-10. Both datasets feature a training set containing 50,000 images and a test set of 10,000 images. SVHN [Netzer et al., 2011], Tiny-ImageNet (TIN) [Le and Yang, 2015] and LSUN (LS) [Yu et al., 2015] in its (c)roped and (r)seized versions, iSUN [Xu et al., 2015], Textures (Tex.) [Cimpoi et al., 2014], Places365 (Places) [Zhou et al., 2017], Gaussian noise (Gauss), and Uniform noise (Unif.) are used as OOD datasets. The ImageNet [Deng et al., 2009] dataset encompasses approximately 1.28 million training examples and 50,000 labeled test instances from 1000 classes. For this large-scale benchmark, in addition to Textures and Places365 (Places), Species [Hendrycks et al., 2022], OpenImage-O (OpenIm) [Wang et al., 2022], iNaturalist (iNat) [Huang and Li, 2021], Sun [Huang and Li, 2021], Semantic Shift Benchmark (SSB) [Vaze et al., 2022], and NINCO [Bitterwolf et al., 2023] are considered.

Evaluation Metrics. Following Narasimhan et al. [2024], we define the evaluation dataset as $S_{\text{all}} \doteq S_{\text{in}} \cup S_{\text{out}}$, a combination of in-d and out-d sets such that $\hat{\pi}_{\text{in}} = |S_{\text{in}}|/\left(|S_{\text{in}}| + |S_{\text{out}}|\right) \approx \pi_{\text{in}}$. From this set, we can compute a few key metrics, such as the empirical coverage, risk, and the area under the risk-coverage curve (AUC-RC or AURC), summarizing the SCOD rejector performance. We also plotted these curves for fine-grained analysis and computed the AUROC for the SCOD problem. We fix $\pi_{\text{in}} = 0.5$ and $c_{\text{fn}} = 0.75$ for all the experiments. All the results are averaged over 10 random seeds.

Empirical Coverage. The empirical coverage counts how many samples are not rejected, i.e.,

$$\hat{\phi}(r) \doteq \frac{1}{|S_{\text{all}}|} \sum_{x \in S_{\text{all}}} \mathbb{1}[r(x) = 0] \approx 1 - b_{\text{rej}}. \tag{11}$$

Empirical SCOD Risk. The empirical SCOD risk, or *joint risk* as in Narasimhan et al. [2024], measures the rate of mistakes of the rejector weighted by c_{fn} on in-d and out-d data. It counts how many misclassified samples are accepted and how many OOD samples are accepted compared to the total amount of accepted samples when a rejector is fixed

Table 2: Comparative analysis of AURC in the black-box SCOD framework between 12 existing OOD detection methods and ours (combining the other 12 methods) for three different models and domains. Results are sorted in descending order by average.

		C-100	SVHN	Text.	Places	Unif.	Avg
VGG-16 (CIFAR-10)	ReAct	0.395	0.357	0.417	0.314	0.259	0.348
	ODIN	0.294	0.248	0.248	0.287	0.181	0.252
	MaxL	0.294	0.248	0.248	0.287	0.180	0.251
	Energy	0.294	0.249	0.248	0.286	0.178	0.251
	Igeood	0.286	0.230	0.256	0.294	0.180	0.249
	KL M	0.280	0.228	0.239	0.285	0.185	0.243
	MSP	0.272	0.221	0.241	0.276	0.182	0.239
	Doctor	0.272	0.221	0.242	0.275	0.181	0.238
	RelMaha	0.274	0.224	0.241	0.253	**0.177**	0.234
	Ours	0.253	0.219	0.241	0.237	0.185	0.227
	Maha	0.244	0.215	0.219	0.241	0.188	0.222
	MCos	0.244	0.209	0.218	0.238	0.182	0.218
	KNN	**0.230**	**0.205**	**0.206**	**0.224**	0.179	**0.208**
		C-10	SVHN	Text.	Places	Unif.	Avg
DenseNet-121 (CIFAR-100)	ReAct	0.514	0.461	0.417	0.427	0.353	0.435
	RMaha	0.367	0.334	0.465	0.364	0.337	0.373
	Maha	0.430	0.361	0.319	0.366	0.294	0.354
	KL M	0.366	0.326	0.363	0.328	0.308	0.338
	Energy	0.355	0.322	0.358	0.323	0.300	0.332
	MaxL	0.353	0.320	0.358	0.324	0.301	0.331
	ODIN	0.353	0.320	0.357	0.323	0.300	0.331
	Doctor	**0.342**	0.322	0.351	0.322	0.309	0.329
	MSP	**0.342**	0.323	0.349	0.321	0.309	0.329
	Ours	0.373	0.318	0.331	0.322	0.288	0.326
	Igeood	0.346	0.315	0.350	**0.318**	0.294	0.325
	KNN	0.374	0.313	0.305	**0.318**	0.276	0.317
	MCos	0.372	**0.310**	**0.297**	0.320	**0.272**	**0.315**
		NINCO	iNat	Text.	Places	OpenIm	Avg
ResNet-101 (ImageNet)	Maha	0.390	0.341	0.310	0.373	0.346	0.352
	Energy	0.343	0.310	0.320	0.318	0.318	0.322
	ODIN	0.341	0.304	0.319	0.315	0.314	0.319
	KNN	0.338	0.312	0.293	0.324	0.321	0.318
	ReAct	0.352	0.304	0.310	0.306	0.314	0.317
	KL M	0.339	0.315	0.304	0.295	0.317	0.314
	MaxL	0.334	0.301	0.304	0.304	0.303	0.309
	RMaha	0.316	0.289	0.303	0.314	0.311	0.307
	Ours	0.331	0.292	0.295	0.303	0.303	0.305
	Doctor	0.312	0.300	0.310	0.296	0.299	0.303
	MSP	**0.306**	**0.281**	0.300	0.291	0.299	0.296
	MCos	0.324	**0.281**	**0.276**	0.296	0.295	0.294
	Igeood	0.312	0.286	0.286	**0.282**	**0.287**	**0.291**

according to a coverage budget. Formally,

$$\hat{R}\left(f,r\right) \doteq \frac{(1-c_{\text{fn}})}{A} \underbrace{\sum_{(x,y)\in S_{\text{in}}} \mathbb{1}[f(x) \neq y, r(x)=0]}_{\text{\# of accepted misclassifications}} + \frac{c_{\text{fn}}}{A} \underbrace{\sum_{x\in S_{\text{out}}} \mathbb{1}[r(x)=0]}_{\text{\# of accepted OOD samples}}, \tag{12}$$

where $A = \sum_{\boldsymbol{x}\in S_{\text{all}}} \mathbb{1}[r(\boldsymbol{x})=0]$ is the total number of accepted samples.

AUC-RC. The area under the risk-coverage curve is computed using the trapezoidal integration rule by discretizing the space of thresholds to compute points of the SCOD risk curve. We considered 10% of the testing points uniformly sampled to be the discretizing thresholds to compute the integral, which accelerates the computations considerably. We report the average results over 10 random seeds, where we observed a standard deviation of less than 10^{-3}.

5 RESULTS AND ANALYSIS

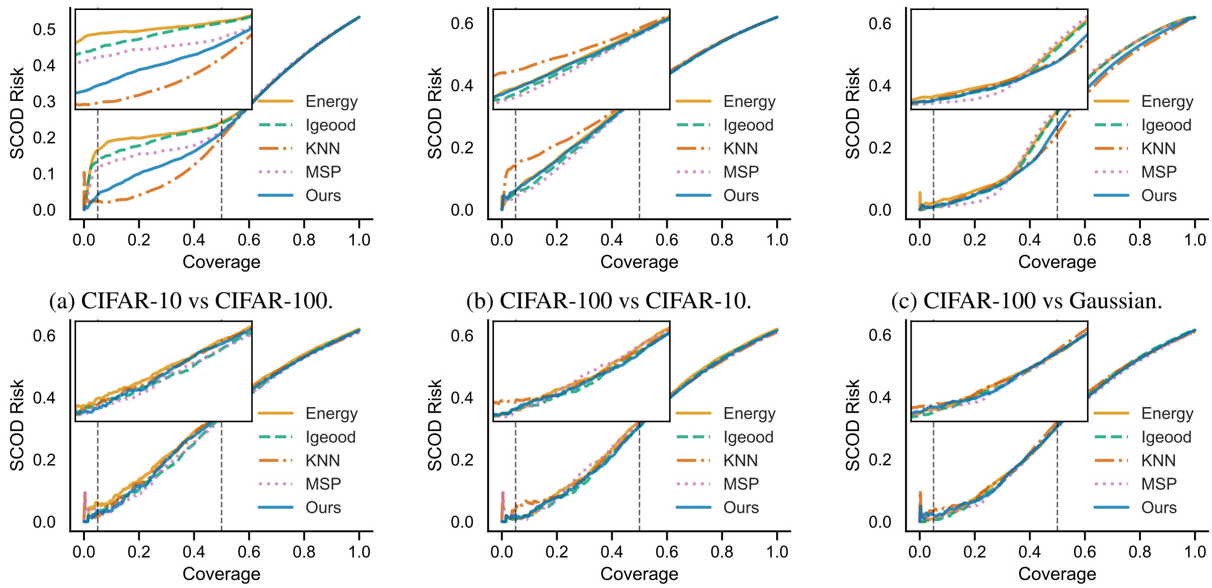
Table 2 showcases the detection performance of the baselines and our method on the SCOD benchmark in terms of AURC. We observe that the proposed solution falls within the top 5 best detection methods on average, but most importantly, it keeps a low distance compared to the best performer for the given task. For CIFAR-10, on average, the worst method degrades performance by 67%, while ours degrades best performance by only 9%. For CIFAR-100, the worst method degrades performance by 38%; in contrast, we are merely 3.5% below the best performer. Finally, for ImageNet, the worst method degrades performance by 21%, while we are just 4.8% below the best method. As a result, we *reduced the worst case risk* by 35%, 25%, and 13% on these three benchmarks, respectively. In addition, we compared our method to baseline aggregation methods in Appendix A.4.

Figure 3 shows the SCOD **risk-coverage curves** for a few tasks on the benchmark. As desired, the proposed method never attains the worst performance, and we show empirically that the performance is much closer to the best method on the specific benchmark than the worst one, confirming the results obtained in Table 2. On ImageNet, results seem more uniform across methods, at least for the benchmark displayed in Figure 3. Extended results are relegated to the Appendix A.6, where you can find all the experimental points, including AURC and AUROC results, used to analyze the results obtained in this paper, which comprises over 20 trained models and further OOD datasets.

We analyzed the **execution time** in Equation (9), as a function of the number of detectors. Notably, the optimization algorithms can be optimized to run in parallel on a GPU, showing no outstanding overhead (cf. Figure 4 in Appendix A.3) with the code available in Listing 1.

Limitations and domain shift scenario. Although the approach introduced in this work guarantees minimal regret and achieves consistent reduction of the worst-case risk across several domains, it is not free from the risk that none of the aggregated detectors is effective against a new given OOD domain. While typically the "robustness to distribution shifts" of OOD detectors is empirically assessed by testing on various benchmarks, *our approach offers the opportunity to establish a formal upper bound on error detection.*

Owing to the probabilistic nature introduced by the PDS framework, we can utilize Ben-David et al. [2010] to obtain



(a) CIFAR-10 vs CIFAR-100. (b) CIFAR-100 vs CIFAR-10. (c) CIFAR-100 vs Gaussian.

(d) ImageNet vs NINCO. (e) ImageNet vs OpenImages. (f) ImageNet vs iNaturalist.

Figure 3: Risk-coverage curves for the SCOD black-box plugin framework for our method and popular OOD detection baselines.

an error bound based on the statistical divergence between the domains where our detector performs well and any new domain. For detailed insights, refer to the proof provided in Appendix A.2. While this might be regarded as a limitation, indicating that no OOD detector can be universally effective across all possible domains Zhang et al. [2021], Fang et al. [2022], it points out that comparing detectors on standard benchmarks often provides a false sense of security.

6 DISCUSSION

In Section 3, we introduce a theoretical framework demonstrating that it is possible to construct a detector that minimizes regret across all potential detectors contained in the set $\mathcal{Q}$.

To recall, in information theory, the concept of minimized regret addresses the challenge of designing a detection method from a set of available options to reduce the risk (i.e., detection error) associated with preemptively choosing a single detector, as is common practice. This approach mitigates the worst-case scenario that arises when selecting a single detector for diverse OOD detection tasks (cf. Figure 1). This key advantage sets our method apart from existing state-of-the-art black-box SCOD plugin methods (cf. Narasimhan et al. [2024]).

We evaluate the proposed framework across a wide range of benchmark datasets (see Sections 4 and 5, and Appendix A.6). While it consistently performs close to the best detector in most cases, every other plugged-in detection method—except ours—inevitably encounters at least one instance where it reaches worst-case detection performance. This finding is supported by the results in Table 2 in Section 4 and Tables 5 to 8 in Appendix A.6.

Additionally, in Appendix A.4, we compare the proposed aggregation method, with less nuanced ones, such as majority voting, and assigning the same weight to all the detectors of the scores. On average, when multiple OOD datasets are considered, our method outperforms these baselines up to two percentage points (cf. right-most colums in Tables 3 and 4), while retaining the theoretical guarantees of minimal regret.

Furthermore, we provide a short analysis of the interpretability features of our method in Appendix A.5, showing in Figure 5 how the scores it assigns differ from those assigned by less nuanced solutions, and how they reflect the nature of the underlying data and the considered detectors.

7 RELATED WORKS

In this section we position our work within the broader context of Selective Classification, Out-of-Distribution detection, and the intersection of these two fields, i.e., Selective Classification with Out-of-Distribution Detection.

7.1 SELECTIVE CLASSIFICATION

A large body of work emphasizes the importance of fitting auxiliary parameters to directly estimate a detection score, aligning with the "learning to reject" paradigm [Chow, 1957, 1970, Geifman and El-Yaniv, 2017, Corbière et al., 2019, Liu et al., 2019, Huang et al., 2020]. Zhu et al. [2023] analyzes how models respond to outliers, aiming to assess the effectiveness of these heuristics in enhancing misclassification performance. In the mathematical framework proposed by Granese et al. [2021], a simple detection method based on the estimated probability of error is introduced.

The investigation by Zhu et al. [2022a] shows that calibration methods often prove counterproductive for failure prediction, offering valuable insights into the underlying reasons. Cen et al. [2023] explores the impact of training settings on misclassification detection performance. Other contributions in this area encompass uncertainty estimation through Bayesian Neural Networks [Gal and Ghahramani, 2016, Lakshminarayanan et al., 2017] and conformal predictions [Gibbs and Candes, 2021]. Zhang et al. [2022] relies on the adaptation to augmented data produced at test time.

7.2 OUT-OF-DISTRIBUTION DETECTION

The taxonomy of post-hoc OOD detection methods delineates three main categories: *confidence-based*, *distance-based*, and *mixed distance-confidence* techniques. Confidence-based methods [Hein et al., 2019, Hendrycks and Gimpel, 2017, Liang et al., 2018, Hsu et al., 2020, Liu et al., 2020, Hendrycks et al., 2022, Sun and Li, 2022], rely on logits or softmax outputs of neural networks. Distance-based methods [Sun et al., 2021, Huang et al., 2021, Zhu et al., 2022b, Colombo et al., 2022, Dong et al., 2021, Dadalto et al., 2022, Song et al., 2022, Lin et al., 2021, Djurisic et al., 2023, Lee et al., 2018, Fort et al., 2021, Sun et al., 2022, Du et al., 2022a, Ming et al., 2023] focus on latent representations by measuring dissimilarities between input samples and training prototypes.

Mixed distance-confidence techniques [Wang et al., 2022, Dadalto et al., 2024, Wu et al., 2023] combine information from both outputs and latent representations. Learning with outlier exposure [Hendrycks et al., 2019, Du et al., 2022b] incorporates outlier samples to regularize shape decision boundaries to be outlier-aware. Benchmarks [Zhang et al., 2023] underscore the absence of a singularly superior method, highlighting the complexity and challenges inherent to OOD detection. The concurrent work [Fan et al., 2024] relies on the idea that an online detector can be trained at test time batches using the linear separability between scores for in-distribution and out-of-distribution data points.

7.3 SELECTIVE CLASSIFICATION WITH OUT-OF-DISTRIBUTION DETECTION

Narasimhan et al. [2024], Katz-Samuels et al. [2022], Xia and Bouganis [2022], simultaneously identifying misclassified samples and samples from outside the training distribution. Xia and Bouganis [2022] empirically observes that softmax-based scores are superior in misclassification on a few benchmarks, and combining it with class agnostic features such as the norm of the output features of the penultimate layer or the residual score introduced in Wang et al. [2022] could improve SCOD detection. Narasimhan et al. [2024] provides a plugin framework to combine off-the-shelf SC and OOD detection methods to achieve a Bayes-optimal black-box rejector, from which this work extends to introduce our combined SCOD rejector that minimizes regret on the target task.

A different approach, distinct from the one presented in this work and based on orthogonal assumptions, is proposed in Franc et al. [2024]. It aligns with the white-box scenarios outlined in Narasimhan et al. [2024], and in analogy with Gomes et al. [2024] aims to define a data-driven Bayesian SCOD detector that requires a training algorithm, in-d, and notably multiple out-d data shots. Katz-Samuels et al. [2022] also leverages learning the optimization of a surrogate constrained optimization problem using unlabeled in-the-wild data, but in a framework that fits the white-box approach in Narasimhan et al. [2024]. The concurrent work [Vishwakarma et al., 2024] incorporates expert human feedback to safely update the OOD detection threshold.

8 CONCLUSION

We proposed a regret minimization-based framework to aggregate several off-the-shelf OOD detection methods with the SCOD paradigm. Crucially, this method is zero-shot, easy to implement, cost-effective, and can be applied to several OOD methods. We show that our framework can consistently reduce the worst-case detection risk across several domains while also attaining performance comparable to SOTA solution when they are plugged in the black-box SCOD framework Narasimhan et al. [2024].

SCOD presents a particularly challenging detection problem in Machine Learning, as it involves detecting both misclassified samples and those originating from outside the training distribution. The novel approach we proposed in this work aligns distance-based and confidence-based detectors, and formally aggregates their decisions, providing a principled way to combine off-the-shelf detectors, and providing a new perspective on the cat-and-mouse game of developing new ones—an endeavor often driven solely by empirical comparisons and lacking theoretical guarantees.

We are hopeful that this work will stimulate further research in the field of SCOD, particularly toward a more rigorous assessment of the worst-case risk of newly proposed methods. By providing a principled baseline through the MRPSA solution, our framework offers a modular and flexible design fostering cumulative progress rather than isolated advancements.

References

Suguru Arimoto. An algorithm for computing the capacity of arbitrary discrete memoryless channels. *IEEE Trans. Inf. Theory*, 18(1):14–20, 1972. doi: 10.1109/TIT.1972.1054753.

A. Barron, J. Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, 1998. doi: 10.1109/18.720554.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010. doi: 10.1007/s10994-009-5152-4.

Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? Fixing imagenet out-of-distribution detection evaluation. In *International Conference on Machine Learning*, 2023.

Jun Cen, Di Luan, Shiwei Zhang, Yixuan Pei, Yingya Zhang, Deli Zhao, Shaojie Shen, and Qifeng Chen. The devil is in the wrongly-classified samples: Towards unified open-set recognition. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

C. K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, EC-6(4):247–254, 1957. doi: 10.1109/TEC.1957.5222035.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Trans. Inf. Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3606–3613, 2014. doi: 10.1109/CVPR.2014.461.

Pierre Colombo, Eduardo Dadalto, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. Beyond mahalanobis distance for textual ood detection. In *Advances in Neural Information Processing Systems*, 2022.

Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, pages 2898–2909, 2019.

Eduardo Dadalto, Florence Alberge, Pierre Duhamel, and Pablo Piantanida. Igeood: An information geometry approach to out-of-distribution detection. In *International Conference on Learning Representations*, 2022.

Eduardo Dadalto, Florence Alberge, Pierre Duhamel, and Pablo Piantanida. Combine and conquer: A meta-analysis on data shift and out-of-distribution detection. *Trans. Mach. Learn. Res.*, 2024, 2024.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *The Eleventh International Conference on Learning Representations*, 2023.

Xin Dong, Junfeng Guo, Ang Li, Wei-Te Mark Ting, Cong Liu, and H. T. Kung. Neural mean discrepancy for efficient out-of-distribution detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19195–19205, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Xuefeng Du, Gabriel Gozum, Yifei Ming, and Yixuan Li. SIREN: Shaping representations for detecting out-of-distribution objects. In *Advances in Neural Information Processing Systems*, 2022a.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Sharon Li. Towards unknown-aware learning with virtual outlier synthesis. In *International Conference on Learning Representations*, 2022b.

Ke Fan, Tong Liu, Xingyu Qiu, Yikai Wang, Lian Huai, Zeyu Shangguan, Shuang Gou, Fengjian Liu, Yuqian Fu, Yanwei Fu, and Xingqun Jiang. Test-time linear out-of-distribution detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02242.

Zhen Fang, Yixuan Li, Jie Lu, Jiahua Dong, Bo Han, and Feng Liu. Is out-of-distribution detection learnable? In *Advances in Neural Information Processing Systems*, 2022.

Leo Feng, Mohamed Osama Ahmed, Hossein Hajimirsadeghi, and Amir H. Abdi. Stop overcomplicating selective classification: Use max-logit. *CoRR*, abs/2206.09034, 2022. doi: 10.48550/arXiv.2206.09034.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2021.

Vojtech Franc, Jakub Paplhám, and Daniel Prruvsa. SCOD: from heuristics to theory. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXIV*, volume 15142 of *Lecture Notes in Computer Science*, pages 424–441. Springer, 2024. doi: 10.1007/978-3-031-72907-2_25.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 1050–1059. JMLR.org, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR, 2019.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pages 1660–1672. Curran Associates, Inc., 2021.

Eduardo Dadalto Câmara Gomes, Marco Romanelli, Georg Pichler, and Pablo Piantanida. A data-driven measure of relative uncertainty for misclassification detection. In *The Twelfth International Conference on Learning Representations*, 2024.

Federica Granese, Marco Romanelli, Daniele Gorla, Catuscia Palamidessi, and Pablo Piantanida. DOCTOR: A simple method for detecting misclassification errors. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 5669–5681, 2021.

Federica Granese, Marco Romanelli, and Pablo Piantanida. Optimal zero-shot detector for multi-armed attacks. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 2467–2475. PMLR, 02–04 May 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 41–50, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Xiaodong Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning*, 2022.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Alexander Shepherd, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist challenge 2017 dataset. *2017 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications (2017). doi: 10.48550, 2017.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10948–10957, 2020.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269, 2017. doi: 10.1109/CVPR.2017.243.

Lang Huang, Chao Zhang, and Hongyang Zhang. Self-adaptive training: beyond empirical risk minimization. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8706–8715, 2021.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 2021.

Julian Katz-Samuels, Julia B. Nakhleh, Robert D. Nowak, and Yixuan Li. Training OOD detectors in their natural habitats. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 10848–10865. PMLR, 2022.

Julian Katz-Samuels, Julia B. Nakhleh, Robert D. Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, 2022.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. 2017.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. 2015.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31*, pages 7167–7177. Curran Associates, Inc., 2018.

Yonghoon Lee and Rina Barber. Distribution-free inference for regression: discrete, continuous, and in between. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 7448–7459, 2021.

Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, Jinfeng Yi, and Bowen Zhou. Trustworthy ai: From principles to practices. *ACM Computing Surveys*, 2023.

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

Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020.

Ziyin Liu, Zhikang Wang, Paul Pu Liang, Ruslan Salakhutdinov, Louis-Philippe Morency, and Masahito Ueda. Deep gamblers: Learning to abstain with portfolio theory. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10622–10632, 2019.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations*, 2023.

Harikrishna Narasimhan, Aditya Krishna Menon, Wittawat Jitkrittum, and Sanjiv Kumar. Plugin estimators for selective classification with out-of-distribution detection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

Georg Pichler, Marco Romanelli, Divya Prakash Manivannan, Prashanth Krishnamurthy, Farshad khorrami, and Siddharth Garg. On the (in)feasibility of ML backdoor detection as an hypothesis testing problem. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4051–4059. PMLR, 02–04 May 2024.

Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with Gram matrices. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8491–8501. PMLR, 13–18 Jul 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Yue Song, Nicu Sebe, and Wei Wang. Rankfeat: Rank-1 feature removal for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, 2022.

Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *European Conference on Computer Vision*, 2022.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *NeurIPS*, 2021.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR, 17–23 Jul 2022.

Engkarat Techapanurak, Masanori Suganuma, and Takayuki Okatani. Hyperparameter-free out-of-distribution detection using cosine similarity. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, November 2020.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. In *International Conference on Learning Representations*, 2022.

Harit Vishwakarma, Heguang Lin, and Ramya Korlakai Vinayak. Taming false positives in out-of-distribution detection with human feedback. In *International Conference on Artificial Intelligence and Statistics, 2-4 May 2024, Palau de Congressos, Valencia, Spain*, volume 238 of *Proceedings of Machine Learning Research*, pages 1486–1494. PMLR, 2024. URL `https://proceedings.mlr.press/v238/vishwakarma24a.html`.

John von Neumann. Zur theorie der gesellschaftsspiele. *Mathematische Annalen*, 100:295–320, 1928.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4911–4920, 2022.

Xinheng Wu, Jie Lu, Zhen Fang, and Guangquan Zhang. Meta OOD learning for continuously adaptive OOD detection. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19296–19307. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01773.

Guoxuan Xia and Christos-Savvas Bouganis. Augmenting softmax information for selective classification with out-of-distribution data. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 1995–2012, December 2022.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R. Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking, 2015.

F. Yu, Y. Zhang, Shuran Song, Ari Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

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. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023.

Lily H. Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12427–12436. PMLR, 2021.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

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*, 2017.

Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Rethinking confidence calibration for failure prediction. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXV*, page 518–536, Berlin, Heidelberg, 2022a. Springer-Verlag. ISBN 978-3-031-19805-2. doi: 10.1007/978-3-031-19806-9_30.

Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-Lin Liu. Openmix: Exploring outlier samples for misclassification detection. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Yao Zhu, YueFeng Chen, Chuanlong Xie, Xiaodan Li, Rong Zhang, Hui Xue', Xiang Tian, bolun zheng, and Yaowu Chen. Boosting out-of-distribution detection with typical features. In *Advances in Neural Information Processing Systems*, 2022b.

Optimal Zero-shot Regret Minimization for Selective Classification With Out-of-Distribution Detection

Eduardo Dadalto [1] **Marco Romanelli**[2]

[1]Helsing, Paris, France
[2]Computer Science Dept., Hofstra University, Hempstead, New York, USA

A APPENDIX

A.1 PROOFS

A.1.1 Proof of Equation (7)

Proof.

$$\begin{aligned}\max_{k\in\mathcal{K}} \mathbb{E}_{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}\left[-\log \mathbb{Q}_{Z|\boldsymbol{x}}\right] &= \max_{k\in\mathcal{K}}\left[\mathbb{E}_{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}\left[-\log \mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}\right] + \mathbb{E}_{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}\left[\log\left(\frac{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}{\mathbb{Q}_{Z|\boldsymbol{x}}}\right)\right]\right]\\ &\leq \max_{k\in\mathcal{K}} \mathbb{E}_{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}\left[-\log \mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}\right] + \max_{k\in\mathcal{K}} \mathbb{E}_{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}\left[\log\left(\frac{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}{\mathbb{Q}_{Z|\boldsymbol{x}}}\right)\right].\end{aligned}$$

□

A.1.2 Proof of Equation (8)

Proof. The equality holds by noticing that

$$\max_{P_\Omega} \mathbb{E}_\Omega\left[D_{\mathrm{KL}}\left(\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}} \| \mathbb{Q}_{Z|\boldsymbol{x}}\right)\right] \leq \max_{k\in\mathcal{K}} \mathbb{E}_{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}\left[\log\left(\frac{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}{\mathbb{Q}_{Z|\boldsymbol{x}}}\right)\right],$$

and moreover,

$$\max_{k\in\mathcal{K}} \mathbb{E}_{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}\left[\log\left(\frac{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}{\mathbb{Q}_{Z|\boldsymbol{x}}}\right)\right] = \mathbb{E}_{\bar{\Omega}}\left[D_{\mathrm{KL}}\left(\mathbb{Q}^{(\bar{\Omega})}_{Z|\boldsymbol{x}} \| \mathbb{Q}_{Z|\boldsymbol{x}}\right)\right],$$

for a uniformly distributed random variable $\bar{\Omega}$ for the set of maximizers $\overline{\mathcal{K}} = \arg\max_{k\in\mathcal{K}} \mathbb{E}_{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}\left[\log\left(\frac{\mathbb{Q}^{(k)}_{Z|\boldsymbol{x}}}{\mathbb{Q}_{Z|\boldsymbol{x}}}\right)\right]$, zero otherwise. □

*Work done while working at Université Paris-Saclay CNRS CentraleSupélec.

A.1.3 Proof of Equation (9)

Proof. Let us consider a zero-sum game with a concave-convex mapping defined on a product of convex sets. The sets of all probability distributions $\mathbb{Q}_{Z|\boldsymbol{x}}$ and P_Ω are two nonempty convex sets, bounded and finite-dimensional. On the other hand, $\left(P_\Omega, \mathbb{Q}_{Z|\boldsymbol{x}}\right) \to \mathbb{E}_\Omega\left[D_{\mathrm{KL}}\left(\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}} \| \mathbb{Q}_{Z|\boldsymbol{x}}\right)\right]$ is a concave-convex mapping, i.e., $P_\Omega \to \mathbb{E}_\Omega\left[D_{\mathrm{KL}}\left(\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}} \| \mathbb{Q}_{Z|\boldsymbol{x}}\right)\right]$ is concave and $\mathbb{Q}_{Z|\boldsymbol{x}} \to \mathbb{E}_\Omega\left[D_{\mathrm{KL}}\left(\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}} \| \mathbb{Q}_{Z|\boldsymbol{x}}\right)\right]$ is convex for every $\left(P_\Omega, \mathbb{Q}_{Z|\boldsymbol{x}}\right)$. Then, by classical min-max theorem von Neumann [1928] we have

$$\min_{\mathbb{Q}_{Z|\boldsymbol{x}}} \max_{P_\Omega} \mathbb{E}_\Omega\left[D_{\mathrm{KL}}\left(\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}} \| \mathbb{Q}_{Z|\boldsymbol{x}}\right)\right] = \max_{P_\Omega} \min_{\widehat{\mathbb{Q}}_{Z|\boldsymbol{x}}} \mathbb{E}_\Omega\left[D_{\mathrm{KL}}\left(\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}} \| \mathbb{Q}_{Z|\boldsymbol{x}}\right)\right].$$

For the next result, it is enough to show that

$$\min_{\widehat{\mathbb{Q}}_{Z|\boldsymbol{x}}} \mathbb{E}_\Omega\left[D_{\mathrm{KL}}\left(\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}} \| \mathbb{Q}_{Z|\boldsymbol{x}}\right)\right] = I_{\boldsymbol{x}}(\Omega; Z), \tag{13}$$

for every random variable Ω distributed according to an arbitrary probability distribution P_Ω and each distribution $\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}}$. We begin by showing that

$$\mathbb{E}_\Omega\left[D_{\mathrm{KL}}\left(\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}} \| \mathbb{Q}_{Z|\boldsymbol{x}}\right)\right] \geq I_{\boldsymbol{x}}(\Omega; Z),$$

for any arbitrary distributions P_Ω and $\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}}$. To this end, we use the following identities:

$$\begin{aligned}
\mathbb{E}_\Omega\left[D_{\mathrm{KL}}\left(\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}} \| \mathbb{Q}_{Z|\boldsymbol{x}}\right)\right] &= \mathbb{E}_\Omega \mathbb{E}_{\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}}}\left(\log \frac{\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}}}{\mathbb{Q}_{Z|\boldsymbol{x}}}\right) \\
&= \mathbb{E}_\Omega \mathbb{E}_{\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}}}\left(\log \frac{\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}}}{P_Z}\right) + D_{\mathrm{KL}}\left(P_Z \| \mathbb{Q}_{Z|\boldsymbol{x}}\right) \\
&= I_{\boldsymbol{x}}(\Omega; Z) + D_{\mathrm{KL}}\left(P_Z \| \mathbb{Q}_{Z|\boldsymbol{x}}\right) \geq I_{\boldsymbol{x}}(\Omega; Z),
\end{aligned} \tag{14}$$

where P_Z represents the marginal distribution of $\mathbb{Q}^{(\Omega)}_{Z|\boldsymbol{x}}$ w.r.t. P_Ω and the last inequality holds for the fact that the KL divergence is non-negative. Finally, it is easy to check that by for $\mathbb{Q}_{Z|\boldsymbol{x}} = P_Z$ the lower bound in (14) holds. As a consequence, this proves the identity in expression (13). By taking the maximum overall probability distributions P_Ω at both sides of expression (13) the claim follows. $\square$

A.2 ON THE CONSEQUENCES OF DOMAIN SHIFT

So far we have described an optimal solution to aggregate $|\mathcal{K}|$ detectors (cf. Equation (10)) such that each of them has is assumed to effectively detect in-d samples and out-d samples drawn from a certain $\mathbb{P}^{(k)}_{\text{out}}$, i.e. the *source domain*. Let us now suppose that a new $k^* \notin \mathcal{K}$ is introduced. Clearly, it may be the case that none of the detectors we aggregated is effectively deployable for this task, thus one may wonder whether the aggregated detector will be able to work on k^*, i.e. the *new domain*. In this section, we consider this problem and provide an upper bound on detection error for the new domain as a function of the detection error for the previous domain.

Let us consider a detector r, like the one defined in Equation (10). Let us also assume a function $f^S : \mathbb{R}^d \to \{0, 1\}$, i.e. the source label function (oracle) which assigns a label to any input sample distributed according to the source domain. Let us define P^S_X, the distribution of the input variable X (in-d or out-d) over the input space $\mathbb{R}^d$, where the out-d samples are generated according to the possible $|\mathcal{K}|$ out-d of which our aggregated detector is aware.

Similarly, we define $f^T : \mathbb{R}^d \to \{0, 1\}$, i.e. the label function relative to the new domain. The new (testing) domain, defined as P^T_X is the distribution of the input X (in-d or out-d), where the OOD samples are generated and indexed with $k^\star$, which are new to our detector.

We can now define the source error:

$$P_e^S(r) \doteq \mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left[\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^S(\boldsymbol{x})\right]\right], \tag{15}$$

and the error on the new domain:

$$P_e^T(r) \doteq \mathbb{E}_{\boldsymbol{x}\sim P_X^T}\left[\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^T(\boldsymbol{x})\right]\right]. \tag{16}$$

Let

$$d\left(P^S_{X|Z=1}, P^T_{X|Z=1}\right) \doteq 2\sup_{B\in\beta}\left|\Pr_S(B)-\Pr_T(B)\right|, \tag{17}$$

where β is the set of measurable subsets under the noise distributions $P^S_{X|Z=1}$, and $P^T_{X|Z=1}$. Then, according to Ben-David et al. [2010],

$$\begin{aligned} P_e^T(r) \leq P_e^S(r) &+ d\left(P^S_{X|Z=1}, P^T_{X|Z=1}\right) \\ &+ \min\left\{\mathbb{E}_{\boldsymbol{x}\sim P_X^S}[|f^S(\boldsymbol{x})-f^T(\boldsymbol{x})|],\right. \\ &\left.\mathbb{E}_{\boldsymbol{x}\sim P_X^T}[|f^S(\boldsymbol{x})-f^T(\boldsymbol{x})|]\right\}. \end{aligned} \tag{18}$$

Intuitively, as the detector has never seen samples from the new domain, it is expected to perform worse on it. Conversely, the above bound indicates that the loss in terms of performance is expected to be low proportionally to a small $d\left(P^S_{X|Z=1}, P^T_{X|Z=1}\right)$ of the noises between the domains. This proof is adapted from Ben-David et al. [2010].

Proof.

$$\begin{aligned} P_e^T(r) &= P_e^T(r)+P_e^S(r)-P_e^S(r)+\mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left[\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^T(\boldsymbol{x})\right]\right] \\ &\quad -\mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left[\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^T(\boldsymbol{x})\right]\right] & (19)\\ &\leq P_e^S(r)+\left|\mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left[\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^T(\boldsymbol{x})\right]\right]-P_e^S(r)\right| + \\ &\quad \left|P_e^T(r)-\mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left[\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^T(\boldsymbol{x})\right]\right]\right| & (20)\\ &\leq P_e^S(r)+\mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left|\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^T(\boldsymbol{x})\right]-\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^S(\boldsymbol{x})\right]\right| + \\ &\quad \left|P_e^T(r)-\mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left[\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^T(\boldsymbol{x})\right]\right]\right| & (21)\\ &\leq P_e^S(r)+\mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left|f^T(\boldsymbol{x})-f^S(\boldsymbol{x})\right|+\left|P_e^T(r)-\mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left[\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^T(\boldsymbol{x})\right]\right]\right| & (22)\\ &\leq P_e^S(r)+\mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left|f^T(\boldsymbol{x})-f^S(\boldsymbol{x})\right|+d\left(P^S_{X|Z=1}, P^T_{X|Z=1}\right). & (23) \end{aligned}$$

Notice that by choosing to add and subtract $\mathbb{E}_{\boldsymbol{x}\sim P_X^T}\left[\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^T(\boldsymbol{x})\right]\right]$ instead of $\mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left[\mathbb{1}\left[r\left(\boldsymbol{x}\right)\neq f^T(\boldsymbol{x})\right]\right]$, we would get the term $\mathbb{E}_{\boldsymbol{x}\sim P_X^T}\left|f^T(\boldsymbol{x})-f^S(\boldsymbol{x})\right|$, instead of $\mathbb{E}_{\boldsymbol{x}\sim P_X^S}\left|f^T(\boldsymbol{x})-f^S(\boldsymbol{x})\right|$. Therefore, the final result holds true:

$$\begin{aligned} P_e^T(r) \leq P_e^S(r) &+ d\left(P^S_{X|Z=1}, P^T_{X|Z=1}\right) \\ &+ \min\left\{\mathbb{E}_{\boldsymbol{x}\sim P_X^S}[|f^S(\boldsymbol{x})-f^T(\boldsymbol{x})|],\right. \\ &\left.\mathbb{E}_{\boldsymbol{x}\sim P_X^T}[|f^S(\boldsymbol{x})-f^T(\boldsymbol{x})|]\right\}. \end{aligned} \tag{24}$$

□

A.3 BLAHUT-ARIMOTO ALGORITHM TIME ANALYSIS

The Blahut-Arimoto algorithm can be accelerated with parallelized computing. Notably, we observe in Figure 4 that the processing times are negligible when implemented on a GPU when compared to the inference time of a deep neural network in classic tasks. Thus, our algorithm does not represent a bottleneck in computation when deployed.

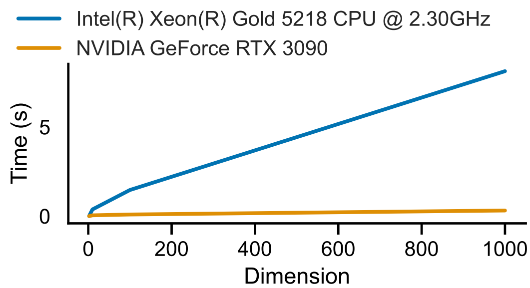


Figure 4: Time analysis for the Blahut-Arimoto algorithm, where dimension is the hypothetical number of detectors to combine.

```python
import torch


def blahut_arimoto(probs: torch.Tensor, max_iter: int = int(1e6), tol: float = 1e-6):
    num_samples, num_detectors, _ = scores.shape
    weights = torch.ones(num_samples, num_detectors, 1) / num_detectors

    for _ in range(max_iter):
        q = torch.mul(weights, probs)
        q = q / torch.sum(q, dim=1, keepdim=True)

        w = torch.prod(torch.pow(q, probs), dim=2, keepdim=True)
        w = w / torch.sum(w, dim=1, keepdim=True)

        tolerance = torch.linalg.norm(w - weights) / torch.linalg.norm(weights)
        weights = w

        if tolerance < tol:
            break

    return weights
```

Listing 1: Blahut-Arimoto algorithm implementation with PyTorch [Paszke et al., 2019].

A.4 COMPARISON WITH BASELINE AGGREGATION ALGORITHMS

We ran experiments with baseline aggregation algorithms and compared their performance to our MRPSA framework on CIFAR-10 and ImageNet benchmarks in Tables 3 and 4, respectively. The *Average* baseline is a simple mean between PDS transformed scores from the same 12 off-the-shelf detectors as in the main experiments in Table 2. The two majority vote methods are based on an evaluation of the detectors' individual decisions on each sample. The aggregation is based on the majority vote, i.e. a weight of 1 is assigned to a detector in agreement with the majority of the detectors, and 0 otherwise. In particular, we consider two variants: in one case we pick a *random* detector within the majority group, in the other case we pick the most *confident* detector within the majority group. In case of a tie, the decision on which detector to pick is made randomly over all the detectors. Even though average and majority vote paradigm's might achieves comparable performance in our benchmarks and others, it is important to stress that these solutions, in stark contrast with ours aggregation, do not guarantee optimality within the regret minimization framework. Thus, despite comparable performance w.r.t. our solution, they do not come with a theoretical guarantee of robustness.

Table 3: Comparative analysis of AURC in the black-box SCOD framework between three baseline aggregation methods (combining 12 off-the-shelf methods) for ResNet-34 trained on CIFAR-10. Results are sorted in descending order by average.

	C-100	SVHN	iSUN	LS (c)	LS (r)	TIN (c)	TIN (r)	Tex.	Places	Unif.	Gauss	Avg
Majority (confident)	0.283	0.206	0.214	0.187	0.205	0.193	0.241	0.299	0.273	0.189	0.222	0.229
Majority (random)	0.272	0.202	0.207	<u>0.182</u>	0.201	0.190	0.231	0.291	0.262	<u>0.183</u>	<u>0.205</u>	0.221
Average	0.254	**0.197**	<u>0.201</u>	<u>0.182</u>	<u>0.195</u>	**0.188**	0.221	0.252	0.248	<u>0.184</u>	0.209	0.212
Ours	**0.242**	0.203	<u>0.201</u>	0.187	<u>0.196</u>	0.196	**0.213**	**0.232**	**0.235**	0.188	<u>0.204</u>	**0.209**

Table 4: Comparative analysis of AURC in the black-box SCOD framework between three baseline aggregation methods (combining 12 off-the-shelf methods) for ResNet-50 trained on ImageNet. Results are sorted in descending order by average.

	iNat.	Species	Places	OpenIm.	SSB (e)	Tex.	NINCO	SSB (h)	Avg
Majority (confident)	0.307	0.330	0.327	0.325	0.321	0.329	0.347	0.388	0.336
Majority (random)	0.295	0.327	0.319	0.322	0.315	0.315	0.342	0.384	0.328
Average	**0.291**	0.319	<u>0.315</u>	0.310	<u>0.308</u>	0.310	**0.332**	**0.374**	<u>0.321</u>
Ours	0.293	**0.315**	<u>0.314</u>	**0.306**	<u>0.308</u>	**0.302**	0.335	0.388	<u>0.321</u>

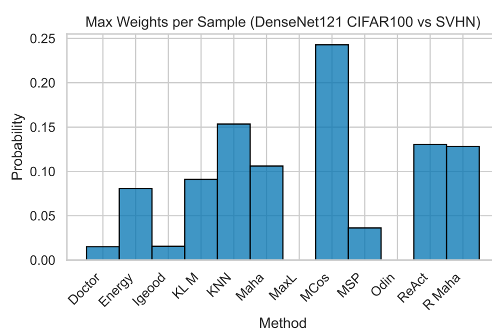
A.5 INTERPRETABILITY OF OUR FRAMEWORK

Figure 5 sheds a light on the interpretability of our method. Due to its inherent design, it aims to perform robust detection across all potential tasks captured by the available pool of detectors. As a result, the method may assign non-zero weights to detectors that are suboptimal for the specific task.

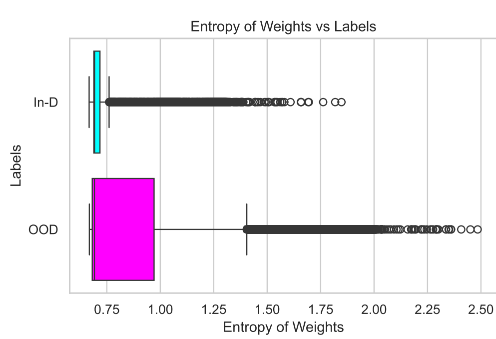
In particular, in Figure 5a we consider a DenseNet121 model trained on CIFAR-100 and focus on the SCOD detection task, with out-of-distribution (OOD) samples from the SVHN dataset. Crucially, the best stand-alone detector is MCos (Maximum Cosine Similarity), and we observe that our aggregation successfully highlights this by assigning it the highest weight—most importantly, without requiring any training or side information about the specific out-distribution used during evaluation.

Figure 5b shows that the weight assignment reflects an intrinsic characteristic of the underlying pool of detectors. These detectors, while trained to recognize specific OOD misclassified samples, are all exposed to the same in-distribution samples once the target DenseNet121 model trained on CIFAR-100 is used. This is evidenced by the low entropy observed for in-distribution samples (blue boxplot), where different detectors tend to agree and can therefore be assigned similar weights. In contrast, entropy is higher for OOD samples (purple boxplot), where detectors are more likely to disagree—necessitating a more nuanced weight assignment. This is an important interpretability feature that is inherent to our proposed solution, that would be lost in less nuanced aggregation methods, such as the average one.

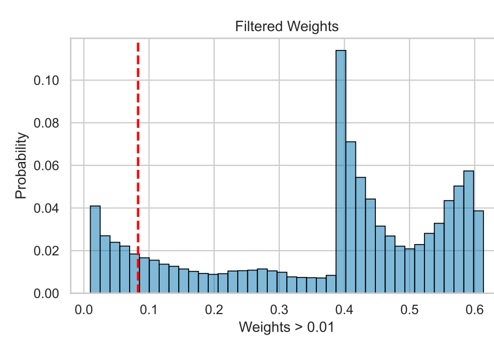
Finally, Figure 5c compares the weights distribution for the average aggregation (dashed line) and our method (blue histogram). The average aggregation assigns the same weight to any of the 12 considered detectors, while our aggregation provides a more nuanced weight assignment, which is able to highlight the best performing detectors, as shown in Figure 5a. Though in terms of hard decision the two methods may exhibit similar performance in some cases, our method provides not only a theoretically grounded approach, but also a more interpretable one, as it allows us to understand the behavior of the detectors and their agreement on the samples.



(a) Average maximum weight histogram attributed to each of the individual detectors with CIFAR100 as in-distribution and SVHN as out-of-distribution.



(b) Shannon Entropy values computed from the weights of our MRPSA framework.



(c) Histogram of individual weight values of the MRPSA framework.

Figure 5: Interpretability plots of MRPSA showcasing interesting properties of the aggregation weights.

A.6 ADDITIONAL RESULTS

Table 5: Comparative analysis of AURC in the black-box SCOD framework between 12 existing OOD detection methods and our method (combining the other 12 methods) for CIFAR-10 models. Results are sorted in descending order by average AURC.

Model	Method	AVG RC	AVG ROC	C-100 RC	C-100 ROC	GAUSS RC	GAUSS ROC	iSUN RC	iSUN ROC	LS (C) RC	LS (C) ROC	LS (R) RC	LS (R) ROC	PLACES RC	PLACES ROC	SVHN RC	SVHN ROC	TEX. RC	TEX. ROC	TIN (C) RC	TIN (C) ROC	TIN (R) RC	TIN (R) ROC	UNIF. RC	UNIF. ROC
DenseNet-121 (CIFAR-10)	RMaha	0.387	0.713	0.386	0.721	0.595	0.374	0.252	0.891	0.520	0.570	0.250	0.902	0.359	0.773	0.586	0.428	0.428	0.704	0.264	0.902	0.287	0.851	0.327	0.724
	ReAct	0.282	0.797	0.314	0.770	0.315	0.690	0.237	0.885	0.273	0.824	0.234	0.889	0.245	0.874	0.280	0.797	0.339	0.709	0.274	0.810	0.270	0.824	0.316	0.692
	KL M	0.254	0.905	0.304	0.852	0.261	0.898	0.252	0.918	0.233	0.921	0.242	0.929	0.302	0.865	0.220	0.931	0.270	0.870	0.247	0.904	0.275	0.895	0.194	0.968
	Energy	0.228	0.918	0.270	0.860	0.220	0.905	0.201	0.956	0.227	0.923	0.198	0.961	0.254	0.890	0.214	0.937	0.307	0.823	0.213	0.937	0.212	0.937	0.191	0.970
	MaxL	0.228	0.919	0.269	0.862	0.220	0.907	0.201	0.956	0.227	0.924	0.198	0.961	0.256	0.888	0.214	0.938	0.306	0.825	0.213	0.938	0.212	0.938	0.190	0.973
	ODIN	0.228	0.919	0.269	0.862	0.219	0.907	0.201	0.956	0.227	0.924	0.198	0.961	0.255	0.889	0.214	0.938	0.305	0.825	0.214	0.937	0.212	0.937	0.189	0.972
	<u>Ours</u>	0.226	0.904	0.252	0.873	0.245	0.846	0.201	0.956	0.248	0.858	0.199	0.959	0.228	0.911	0.253	0.837	0.234	0.887	0.203	0.948	0.218	0.930	0.200	0.944
	IGEOOD	0.217	0.929	0.248	0.884	0.221	0.905	0.199	0.958	0.214	0.935	0.196	0.962	0.239	0.904	0.205	0.949	0.262	0.863	0.208	0.940	0.210	0.940	0.186	0.979
	MSP	0.215	0.928	0.243	0.890	0.207	0.931	0.204	0.945	0.217	0.929	0.200	0.951	0.233	0.906	0.205	0.944	0.249	0.883	0.211	0.931	0.212	0.931	0.189	0.972
	Maha	0.215	0.925	0.270	0.832	0.186	0.981	0.224	0.904	0.211	0.929	0.220	0.912	0.237	0.890	0.197	0.957	0.201	0.953	0.205	0.941	0.237	0.884	0.180	0.988
	Doctor	0.215	0.930	0.243	0.891	0.207	0.931	0.204	0.946	0.216	0.930	0.200	0.952	0.232	0.908	0.205	0.945	0.249	0.884	0.211	0.932	0.211	0.933	0.188	0.974
	MCos	0.201	0.954	0.241	0.895	0.194	0.962	0.200	0.955	0.196	0.961	0.196	0.962	0.221	0.926	0.188	0.977	0.197	0.960	0.193	0.967	0.210	0.937	0.177	0.993
	KNN	0.192	0.967	0.227	0.912	0.185	0.980	0.191	0.970	0.188	0.975	0.189	0.974	0.209	0.943	0.184	0.983	0.191	0.969	0.185	0.981	0.198	0.956	0.169	0.994
ResNet-18 (CIFAR-10)	Maha	0.238	0.882	0.263	0.852	0.181	0.982	0.235	0.880	0.229	0.882	0.233	0.885	0.262	0.848	0.302	0.795	0.232	0.894	0.254	0.847	0.239	0.875	0.188	0.966
	KL M	0.229	0.924	0.271	0.875	0.196	0.953	0.226	0.928	0.191	0.972	0.222	0.933	0.283	0.876	0.242	0.898	0.244	0.896	0.215	0.948	0.240	0.911	0.187	0.970
	ReAct	0.214	0.935	0.254	0.883	0.185	0.971	0.206	0.946	0.180	0.983	0.200	0.954	0.238	0.902	0.249	0.881	0.245	0.894	0.193	0.965	0.222	0.926	0.183	0.976
	MaxL	0.205	0.944	0.239	0.897	0.203	0.936	0.193	0.961	0.177	0.989	0.191	0.965	0.228	0.915	0.210	0.931	0.245	0.896	0.183	0.979	0.202	0.946	0.189	0.966
	ODIN	0.205	0.944	0.239	0.897	0.203	0.936	0.193	0.960	0.177	0.989	0.190	0.965	0.227	0.914	0.210	0.931	0.244	0.896	0.183	0.979	0.202	0.946	0.189	0.965
	Energy	0.205	0.944	0.239	0.897	0.204	0.934	0.193	0.961	0.176	0.990	0.190	0.966	0.227	0.914	0.210	0.932	0.244	0.896	0.182	0.980	0.202	0.947	0.189	0.963
	RMaha	0.205	0.938	0.231	0.900	0.196	0.945	0.199	0.947	0.180	0.982	0.196	0.952	0.228	0.909	0.219	0.911	0.219	0.910	0.192	0.963	0.205	0.937	0.187	0.965
	<u>Ours</u>	0.204	0.939	0.230	0.902	0.182	0.977	0.200	0.944	0.188	0.965	0.198	0.949	0.232	0.904	0.221	0.902	0.212	0.923	0.198	0.947	0.206	0.934	0.180	0.981
	MSP	0.204	0.941	0.228	0.904	0.191	0.957	0.199	0.945	0.182	0.978	0.197	0.950	0.227	0.908	0.208	0.928	0.228	0.909	0.190	0.963	0.205	0.935	0.184	0.972
	MCos	0.203	0.942	0.226	0.907	0.183	0.976	0.200	0.946	0.192	0.961	0.197	0.952	0.226	0.913	0.208	0.924	0.213	0.925	0.203	0.941	0.209	0.930	0.180	0.982
	Doctor	0.203	0.941	0.228	0.904	0.191	0.957	0.199	0.946	0.181	0.979	0.197	0.950	0.225	0.909	0.208	0.928	0.228	0.909	0.189	0.964	0.205	0.936	0.184	0.973
	IGEOOD	0.201	0.948	0.229	0.905	0.195	0.953	0.193	0.961	0.177	0.989	0.190	0.966	0.226	0.915	0.208	0.934	0.230	0.909	0.183	0.979	0.201	0.947	0.184	0.975
	KNN	0.198	0.950	0.218	0.917	0.186	0.967	0.191	0.961	0.183	0.977	0.189	0.966	0.219	0.923	0.208	0.928	0.205	0.936	0.194	0.958	0.198	0.948	0.183	0.975
ResNet-34 (CIFAR-10)	RMaha	0.297	0.860	0.342	0.811	0.343	0.786	0.290	0.873	0.222	0.940	0.271	0.892	0.321	0.837	0.285	0.881	0.372	0.789	0.258	0.906	0.312	0.844	0.252	0.897
	ODIN	0.242	0.908	0.308	0.839	0.266	0.851	0.223	0.928	0.180	0.984	0.211	0.943	0.294	0.861	0.208	0.949	0.333	0.826	0.186	0.974	0.260	0.887	0.198	0.952
	MaxL	0.242	0.909	0.308	0.839	0.265	0.851	0.222	0.928	0.180	0.984	0.211	0.943	0.294	0.861	0.208	0.949	0.333	0.826	0.186	0.974	0.260	0.887	0.197	0.952
	Energy	0.242	0.908	0.308	0.839	0.265	0.850	0.222	0.929	0.179	0.984	0.211	0.943	0.295	0.860	0.208	0.949	0.333	0.826	0.186	0.975	0.259	0.887	0.197	0.951
	IGEOOD	0.234	0.917	0.291	0.852	0.257	0.872	0.220	0.933	0.179	0.985	0.208	0.946	0.293	0.864	0.201	0.955	0.299	0.851	0.184	0.975	0.254	0.895	0.190	0.965
	ReAct	0.231	0.922	0.300	0.848	0.204	0.939	0.216	0.936	0.189	0.974	0.206	0.949	0.250	0.894	0.223	0.933	0.315	0.841	0.198	0.960	0.248	0.899	0.187	0.970
	MSP	0.225	0.924	0.275	0.869	0.213	0.922	0.213	0.934	0.185	0.972	0.205	0.944	0.274	0.877	0.199	0.952	0.295	0.862	0.191	0.960	0.238	0.908	0.186	0.967
	Doctor	0.225	0.925	0.275	0.870	0.213	0.922	0.213	0.934	0.185	0.972	0.205	0.944	0.271	0.879	0.200	0.952	0.295	0.863	0.191	0.961	0.237	0.909	0.186	0.968
	KL M	0.212	0.931	0.245	0.885	0.206	0.929	0.209	0.935	0.188	0.968	0.203	0.944	0.246	0.889	0.199	0.950	0.227	0.900	0.203	0.951	0.216	0.921	0.187	0.966
	Maha	0.210	0.929	0.241	0.885	0.186	0.965	0.197	0.947	0.208	0.927	0.196	0.952	0.227	0.904	0.220	0.910	0.212	0.925	0.222	0.913	0.207	0.933	0.192	0.955
	<u>Ours</u>	0.209	0.932	0.242	0.885	0.204	0.929	0.201	0.944	0.187	0.969	0.196	0.952	0.235	0.900	0.203	0.943	0.232	0.892	0.196	0.954	0.213	0.924	0.188	0.964
	MCos	0.197	0.949	0.219	0.914	0.186	0.964	0.191	0.958	0.192	0.957	0.188	0.964	0.217	0.922	0.193	0.954	0.204	0.936	0.199	0.946	0.200	0.944	0.178	0.984
	KNN	0.193	0.955	0.214	0.921	0.187	0.962	0.187	0.965	0.187	0.966	0.185	0.970	0.212	0.929	0.190	0.960	0.198	0.944	0.193	0.955	0.195	0.951	0.177	0.984
VGG-16 (CIFAR-10)	ReAct	0.334	0.732	0.395	0.654	0.363	0.648	0.300	0.794	0.287	0.797	0.300	0.793	0.314	0.759	0.357	0.665	0.417	0.643	0.322	0.758	0.354	0.726	0.260	0.814
	IGEOOD	0.234	0.925	0.286	0.855	0.190	0.979	0.217	0.945	0.218	0.949	0.218	0.945	0.293	0.859	0.231	0.918	0.255	0.898	0.250	0.914	0.235	0.926	0.180	0.991
	ODIN	0.234	0.925	0.294	0.849	0.190	0.978	0.216	0.946	0.215	0.950	0.217	0.945	0.287	0.863	0.248	0.905	0.249	0.903	0.246	0.916	0.230	0.928	0.181	0.991
	MaxL	0.233	0.925	0.294	0.849	0.190	0.978	0.216	0.946	0.215	0.950	0.217	0.945	0.286	0.863	0.249	0.904	0.248	0.903	0.245	0.916	0.230	0.928	0.179	0.991
	Energy	0.233	0.926	0.294	0.849	0.190	0.979	0.216	0.947	0.215	0.951	0.217	0.946	0.286	0.865	0.249	0.904	0.248	0.903	0.246	0.917	0.230	0.929	0.178	0.991
	KL M	0.230	0.922	0.280	0.860	0.197	0.961	0.225	0.931	0.208	0.947	0.226	0.930	0.284	0.857	0.228	0.919	0.239	0.906	0.224	0.925	0.233	0.916	0.185	0.988
	MSP	0.228	0.926	0.272	0.868	0.194	0.964	0.215	0.940	0.215	0.942	0.217	0.939	0.277	0.868	0.221	0.929	0.242	0.905	0.239	0.917	0.228	0.923	0.183	0.989
	Doctor	0.227	0.927	0.272	0.868	0.194	0.966	0.215	0.941	0.214	0.944	0.216	0.940	0.274	0.872	0.221	0.929	0.242	0.906	0.239	0.917	0.227	0.924	0.181	0.991
	RMaha	0.224	0.935	0.274	0.873	0.191	0.975	0.211	0.948	0.212	0.953	0.213	0.947	0.255	0.897	0.225	0.927	0.241	0.916	0.235	0.929	0.226	0.930	0.178	0.992
	<u>Ours</u>	0.217	0.933	0.253	0.878	0.204	0.944	0.204	0.953	0.200	0.960	0.204	0.953	0.236	0.908	0.219	0.924	0.241	0.900	0.215	0.937	0.221	0.927	0.186	0.983
	Maha	0.212	0.936	0.244	0.890	0.193	0.967	0.206	0.948	0.203	0.947	0.203	0.951	0.241	0.900	0.216	0.923	0.219	0.921	0.211	0.936	0.212	0.935	0.188	0.976
	MCos	0.209	0.944	0.244	0.895	0.192	0.970	0.200	0.957	0.200	0.955	0.200	0.957	0.238	0.905	0.209	0.938	0.218	0.928	0.207	0.945	0.207	0.945	0.183	0.990
	KNN	0.201	0.955	0.230	0.910	0.186	0.982	0.194	0.966	0.194	0.967	0.194	0.967	0.224	0.922	0.205	0.945	0.206	0.944	0.201	0.955	0.200	0.956	0.179	0.993
ViT-B/16 (CIFAR-10)	KL M	0.179	0.975	0.185	0.970	0.154	0.997	0.190	0.966	0.174	0.981	0.191	0.965	0.194	0.964	0.166	0.986	0.153	0.997	0.198	0.958	0.213	0.944	0.155	0.997
	RMaha	0.165	0.982	0.167	0.980	0.161	0.981	0.164	0.985	0.163	0.987	0.165	0.983	0.168	0.978	0.163	0.986	0.161	0.988	0.168	0.978	0.171	0.974	0.162	0.981
	MSP	0.163	0.987	0.167	0.983	0.154	0.998	0.166	0.984	0.162	0.989	0.167	0.982	0.163	0.985	0.161	0.991	0.150	0.998	0.169	0.979	0.175	0.971	0.155	0.998
	Doctor	0.160	0.988	0.167	0.983	0.137	0.999	0.166	0.984	0.162	0.990	0.167	0.983	0.163	0.985	0.161	0.991	0.150	0.998	0.169	0.979	0.175	0.972	0.143	0.999
	IGEOOD	0.159	0.989	0.168	0.983	0.143	0.999	0.166	0.984	0.159	0.993	0.166	0.984	0.161	0.988	0.155	0.996	0.147	0.999	0.168	0.982	0.176	0.971	0.143	0.999
	<u>Ours</u>	0.158	0.990	0.164	0.987	0.138	0.997	0.162	0.989	0.159	0.993	0.162	0.989	0.162	0.987	0.158	0.996	0.154	0.998	0.166	0.984	0.171	0.977	0.141	0.997
	ODIN	0.157	0.990	0.166	0.985	0.133	0.999	0.165	0.986	0.158	0.993	0.166	0.985	0.160	0.989	0.155	0.997	0.149	0.999	0.167	0.983	0.175	0.973	0.135	0.999
	MaxL	0.155	0.990	0.166	0.984	0.130	0.999	0.165	0.986	0.159	0.994	0.166	0.985	0.155	0.990	0.156	0.997	0.138	0.999	0.167	0.983	0.175	0.973	0.131	0.999
	Energy	0.155	0.990	0.166	0.984	0.129	0.999	0.165	0.986	0.159	0.994	0.166	0.985	0.155	0.990	0.157	0.997	0.138	0.999	0.167	0.983	0.175	0.974	0.130	0.999
	Maha	0.155	0.992	0.162	0.988	0.133	0.999	0.159	0.993	0.161	0.990	0.159	0.993	0.157	0.993	0.160	0.991	0.149	0.999	0.163	0.987	0.167	0.981	0.135	0.999
	MCos	0.154	0.988	0.164	0.983	0.129	0.999	0.163	0.985	0.159	0.991	0.163	0.986	0.153	0.989	0.157	0.995	0.136	0.999	0.169	0.978	0.177	0.967	0.130	0.999
	ReAct	0.154	0.991	0.164	0.986	0.130	0.999	0.163	0.988	0.158	0.994	0.164	0.987	0.155	0.991	0.157	0.996	0.138	0.999	0.166	0.984	0.173	0.975	0.130	0.999
	KNN	0.153	0.994	0.161	0.990	0.133	0.999	0.157	0.995	0.155	0.997	0.157	0.995	0.159	0.991	0.150	0.998	0.149	0.999	0.161	0.990	0.164	0.985	0.135	0.999

Table 6: Comparative analysis of AURC in the black-box SCOD framework between 12 existing OOD detection methods and our method (combining the other 12 methods) for CIFAR-100. Results are sorted in descending order by average AURC.

		Avg		C-10		Gauss		iSUN		LS (c)		LS (r)		Places		SVHN		Tex.		TIN (c)		TIN (r)		Unif.	
		RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC
DenseNet-121 (CIFAR-100)	ReAct	0.409	0.755	0.513	0.596	0.360	0.825	0.405	0.761	0.411	0.750	0.404	0.768	0.425	0.729	0.460	0.651	0.417	0.746	0.343	0.880	0.407	0.757	0.353	0.841
	RMaha	0.375	0.801	0.367	0.826	0.297	0.920	0.406	0.752	0.371	0.805	0.396	0.765	0.364	0.819	0.334	0.861	0.464	0.703	0.377	0.787	0.411	0.745	0.337	0.831
	Maha	0.341	0.860	0.429	0.737	0.294	0.939	0.346	0.847	0.350	0.840	0.342	0.851	0.366	0.812	0.362	0.816	0.318	0.896	0.311	0.910	0.336	0.863	0.293	0.944
	KL M	0.329	0.883	0.366	0.848	0.295	0.917	0.338	0.872	0.322	0.896	0.335	0.875	0.328	0.885	0.326	0.882	0.362	0.853	0.299	0.930	0.337	0.870	0.308	0.888
	Energy	0.325	0.885	0.355	0.842	0.307	0.902	0.329	0.880	0.316	0.900	0.325	0.887	0.327	0.889	0.321	0.889	0.360	0.841	0.300	0.926	0.336	0.868	0.301	0.916
	ODIN	0.324	0.888	0.353	0.847	0.305	0.905	0.328	0.882	0.314	0.902	0.324	0.888	0.324	0.892	0.320	0.890	0.358	0.845	0.298	0.930	0.335	0.871	0.302	0.913
	MaxL	0.324	0.888	0.353	0.847	0.305	0.906	0.328	0.882	0.314	0.902	0.324	0.888	0.324	0.892	0.320	0.891	0.358	0.846	0.298	0.930	0.335	0.871	0.301	0.914
	MSP	0.323	0.880	0.342	0.856	0.302	0.900	0.332	0.871	0.315	0.895	0.330	0.873	0.322	0.886	0.324	0.878	0.351	0.847	0.295	0.928	0.334	0.866	0.311	0.879
	Doctor	0.322	0.883	0.342	0.858	0.301	0.900	0.331	0.874	0.313	0.899	0.328	0.877	0.320	0.889	0.323	0.881	0.350	0.849	0.293	0.933	0.333	0.869	0.308	0.885
	IGEOOD	0.317	0.894	0.346	0.854	0.300	0.910	0.322	0.888	0.310	0.907	0.318	0.894	0.318	0.897	0.316	0.894	0.350	0.854	0.292	0.935	0.326	0.881	0.295	0.923
	<u>Ours</u>	0.317	0.898	0.375	0.814	0.284	0.957	0.319	0.891	0.315	0.901	0.316	0.897	0.322	0.892	0.321	0.886	0.330	0.872	0.298	0.927	0.320	0.891	0.287	0.951
	KNN	0.310	0.912	0.374	0.828	0.279	0.963	0.317	0.899	0.313	0.907	0.311	0.908	0.320	0.895	0.313	0.901	0.307	0.915	0.285	0.952	0.314	0.901	0.277	0.963
	MCos	0.306	0.919	0.372	0.831	0.278	0.963	0.313	0.908	0.307	0.918	0.309	0.912	0.319	0.899	0.310	0.909	0.296	0.934	0.284	0.954	0.308	0.913	0.271	0.964
ResNet-18 (CIFAR-100)	ReAct	0.367	0.785	0.338	0.840	0.431	0.630	0.357	0.812	0.342	0.821	0.357	0.812	0.363	0.806	0.397	0.756	0.353	0.820	0.364	0.797	0.356	0.813	0.380	0.724
	KL M	0.326	0.854	0.313	0.877	0.384	0.698	0.317	0.880	0.313	0.878	0.316	0.882	0.340	0.854	0.334	0.864	0.332	0.860	0.293	0.907	0.313	0.889	0.329	0.807
	RMaha	0.325	0.838	0.309	0.872	0.390	0.676	0.314	0.865	0.309	0.868	0.312	0.866	0.334	0.854	0.329	0.860	0.325	0.848	0.298	0.883	0.310	0.870	0.347	0.762
	Energy	0.321	0.852	0.303	0.885	0.386	0.696	0.305	0.887	0.310	0.867	0.303	0.888	0.337	0.849	0.323	0.862	0.326	0.858	0.295	0.898	0.299	0.897	0.340	0.784
	MaxL	0.320	0.853	0.303	0.885	0.386	0.695	0.305	0.886	0.307	0.873	0.304	0.887	0.336	0.852	0.322	0.864	0.325	0.860	0.293	0.902	0.299	0.896	0.339	0.787
	ODIN	0.320	0.853	0.303	0.885	0.385	0.694	0.305	0.886	0.308	0.872	0.304	0.887	0.334	0.852	0.322	0.864	0.324	0.859	0.293	0.902	0.299	0.896	0.338	0.786
	IGEOOD	0.320	0.854	0.303	0.885	0.383	0.701	0.305	0.887	0.308	0.871	0.304	0.888	0.336	0.851	0.322	0.863	0.326	0.859	0.293	0.901	0.299	0.897	0.337	0.789
	MSP	0.320	0.854	0.306	0.880	0.386	0.693	0.309	0.880	0.305	0.878	0.307	0.882	0.333	0.856	0.320	0.867	0.324	0.860	0.291	0.906	0.303	0.889	0.334	0.797
	Doctor	0.319	0.854	0.305	0.882	0.385	0.694	0.308	0.881	0.305	0.878	0.306	0.883	0.332	0.857	0.320	0.867	0.324	0.861	0.291	0.906	0.302	0.891	0.334	0.797
	<u>Ours</u>	0.308	0.879	0.300	0.893	0.319	0.823	0.312	0.883	0.294	0.899	0.309	0.886	0.328	0.864	0.330	0.851	0.306	0.883	0.295	0.902	0.310	0.886	0.288	0.894
	MCos	0.308	0.878	0.304	0.884	0.321	0.816	0.306	0.882	0.301	0.894	0.301	0.890	0.332	0.859	0.334	0.860	0.313	0.876	0.281	0.925	0.304	0.887	0.289	0.890
	KNN	0.306	0.880	0.302	0.887	0.317	0.824	0.304	0.886	0.302	0.893	0.298	0.893	0.332	0.856	0.335	0.857	0.315	0.875	0.279	0.926	0.301	0.891	0.287	0.895
	Maha	0.305	0.891	0.314	0.871	0.280	0.917	0.310	0.877	0.302	0.905	0.304	0.886	0.338	0.853	0.346	0.845	0.310	0.885	0.282	0.926	0.310	0.877	0.262	0.960
ResNet-34 (CIFAR-100)	ReAct	0.411	0.717	0.344	0.834	0.532	0.468	0.400	0.740	0.352	0.826	0.398	0.742	0.402	0.737	0.455	0.662	0.384	0.777	0.409	0.739	0.418	0.714	0.429	0.648
	Energy	0.328	0.856	0.305	0.887	0.426	0.672	0.300	0.896	0.327	0.866	0.295	0.904	0.347	0.851	0.342	0.848	0.346	0.856	0.297	0.908	0.305	0.894	0.320	0.831
	ODIN	0.328	0.855	0.305	0.887	0.426	0.670	0.301	0.893	0.326	0.867	0.296	0.901	0.346	0.852	0.341	0.849	0.345	0.856	0.296	0.908	0.305	0.891	0.320	0.828
	MaxL	0.328	0.855	0.305	0.887	0.426	0.670	0.301	0.893	0.326	0.867	0.295	0.901	0.346	0.852	0.341	0.849	0.345	0.856	0.296	0.909	0.305	0.891	0.321	0.827
	IGEOOD	0.326	0.856	0.303	0.888	0.419	0.675	0.300	0.894	0.323	0.870	0.295	0.902	0.344	0.852	0.340	0.850	0.345	0.856	0.294	0.911	0.305	0.893	0.320	0.828
	MSP	0.326	0.852	0.306	0.882	0.409	0.678	0.305	0.881	0.317	0.874	0.301	0.887	0.338	0.857	0.335	0.855	0.344	0.854	0.293	0.910	0.310	0.880	0.325	0.810
	Doctor	0.325	0.853	0.305	0.883	0.408	0.682	0.305	0.882	0.317	0.874	0.301	0.889	0.339	0.858	0.334	0.855	0.344	0.856	0.293	0.911	0.310	0.881	0.324	0.813
	KL M	0.323	0.854	0.307	0.881	0.391	0.692	0.306	0.882	0.322	0.872	0.299	0.888	0.333	0.858	0.344	0.853	0.332	0.858	0.290	0.911	0.307	0.882	0.322	0.814
	<u>Ours</u>	0.316	0.865	0.289	0.913	0.367	0.729	0.309	0.881	0.294	0.906	0.305	0.885	0.328	0.865	0.342	0.838	0.310	0.882	0.302	0.895	0.317	0.871	0.310	0.848
	RMaha	0.315	0.854	0.300	0.883	0.367	0.714	0.305	0.870	0.302	0.890	0.302	0.875	0.324	0.867	0.329	0.862	0.320	0.856	0.285	0.910	0.305	0.871	0.327	0.798
	MCos	0.306	0.877	0.296	0.893	0.328	0.790	0.299	0.886	0.302	0.893	0.296	0.891	0.322	0.870	0.336	0.863	0.312	0.878	0.278	0.927	0.298	0.889	0.297	0.862
	KNN	0.304	0.879	0.294	0.896	0.327	0.792	0.298	0.890	0.300	0.894	0.294	0.895	0.321	0.870	0.331	0.864	0.314	0.878	0.276	0.929	0.297	0.892	0.296	0.864
	Maha	0.302	0.884	0.294	0.897	0.307	0.834	0.299	0.885	0.298	0.898	0.296	0.889	0.321	0.871	0.335	0.863	0.306	0.885	0.278	0.925	0.299	0.886	0.283	0.894
VGG-16 (CIFAR-100)	Energy	0.453	0.746	0.406	0.818	0.803	0.252	0.372	0.854	0.354	0.884	0.366	0.855	0.422	0.808	0.379	0.832	0.392	0.836	0.347	0.892	0.373	0.850	0.764	0.325
	MaxL	0.452	0.746	0.406	0.819	0.803	0.251	0.372	0.853	0.354	0.883	0.366	0.854	0.419	0.810	0.379	0.832	0.391	0.836	0.347	0.891	0.373	0.849	0.763	0.324
	ODIN	0.452	0.746	0.406	0.819	0.801	0.252	0.372	0.852	0.354	0.883	0.366	0.854	0.419	0.810	0.379	0.832	0.390	0.836	0.347	0.891	0.373	0.848	0.763	0.324
	ReAct	0.451	0.740	0.384	0.826	0.649	0.369	0.417	0.801	0.367	0.864	0.421	0.795	0.389	0.830	0.553	0.694	0.404	0.830	0.358	0.873	0.425	0.795	0.588	0.467
	IGEOOD	0.441	0.756	0.394	0.830	0.792	0.267	0.363	0.859	0.346	0.888	0.360	0.859	0.402	0.825	0.370	0.840	0.378	0.847	0.339	0.897	0.365	0.854	0.745	0.351
	MSP	0.432	0.760	0.379	0.839	0.753	0.315	0.367	0.844	0.341	0.883	0.368	0.840	0.387	0.833	0.369	0.832	0.370	0.845	0.336	0.891	0.373	0.838	0.704	0.400
	Doctor	0.431	0.761	0.379	0.840	0.753	0.317	0.367	0.845	0.341	0.885	0.368	0.841	0.387	0.834	0.369	0.833	0.370	0.846	0.335	0.893	0.373	0.839	0.704	0.401
	<u>Ours</u>	0.391	0.793	0.358	0.851	0.567	0.464	0.355	0.857	0.336	0.890	0.354	0.854	0.372	0.846	0.415	0.795	0.355	0.858	0.322	0.902	0.358	0.852	0.514	0.555
	KL M	0.367	0.816	0.374	0.841	0.419	0.650	0.357	0.846	0.336	0.883	0.361	0.841	0.375	0.837	0.357	0.835	0.363	0.848	0.324	0.894	0.361	0.841	0.414	0.662
	RMaha	0.360	0.824	0.360	0.855	0.468	0.586	0.341	0.863	0.321	0.904	0.342	0.858	0.354	0.861	0.349	0.845	0.341	0.870	0.313	0.907	0.344	0.858	0.426	0.657
	MCos	0.358	0.828	0.361	0.856	0.459	0.599	0.337	0.870	0.329	0.887	0.337	0.868	0.356	0.856	0.342	0.857	0.343	0.869	0.313	0.908	0.338	0.868	0.419	0.669
	Maha	0.356	0.831	0.360	0.858	0.438	0.631	0.341	0.865	0.332	0.879	0.342	0.861	0.357	0.855	0.345	0.853	0.343	0.870	0.315	0.901	0.342	0.864	0.395	0.709
	KNN	0.353	0.833	0.358	0.857	0.447	0.614	0.332	0.877	0.325	0.891	0.332	0.876	0.353	0.856	0.340	0.861	0.340	0.871	0.310	0.912	0.334	0.874	0.415	0.673
ViT-B/16 (CIFAR-100)	KL M	0.226	0.932	0.243	0.931	0.185	0.987	0.251	0.897	0.230	0.918	0.254	0.897	0.242	0.904	0.222	0.924	0.201	0.967	0.237	0.918	0.245	0.911	0.181	0.992
	MSP	0.213	0.941	0.212	0.946	0.182	0.992	0.236	0.910	0.225	0.926	0.229	0.915	0.227	0.918	0.226	0.917	0.193	0.973	0.220	0.931	0.223	0.927	0.170	0.994
	Doctor	0.212	0.942	0.212	0.947	0.179	0.993	0.235	0.911	0.225	0.926	0.229	0.916	0.226	0.919	0.226	0.918	0.192	0.974	0.219	0.933	0.222	0.928	0.164	0.994
	MaxL	0.211	0.950	0.209	0.957	0.172	0.990	0.239	0.916	0.237	0.929	0.226	0.928	0.214	0.946	0.221	0.937	0.189	0.982	0.221	0.940	0.224	0.936	0.167	0.990
	ODIN	0.210	0.950	0.209	0.957	0.170	0.990	0.239	0.916	0.237	0.929	0.227	0.926	0.214	0.945	0.221	0.937	0.188	0.982	0.220	0.940	0.223	0.936	0.164	0.990
	Energy	0.210	0.950	0.209	0.957	0.170	0.990	0.239	0.916	0.238	0.929	0.227	0.927	0.214	0.947	0.220	0.938	0.188	0.982	0.221	0.940	0.224	0.936	0.162	0.990
	ReAct	0.210	0.951	0.208	0.957	0.170	0.989	0.237	0.917	0.237	0.930	0.225	0.929	0.212	0.949	0.218	0.940	0.188	0.982	0.221	0.939	0.223	0.936	0.166	0.989
	RMaha	0.208	0.941	0.205	0.948	0.192	0.963	0.218	0.923	0.210	0.936	0.215	0.930	0.222	0.911	0.208	0.944	0.200	0.953	0.210	0.940	0.212	0.938	0.193	0.960
	IGEOOD	0.207	0.952	0.209	0.956	0.173	0.992	0.235	0.920	0.226	0.936	0.226	0.927	0.215	0.943	0.216	0.940	0.186	0.983	0.214	0.945	0.218	0.940	0.162	0.992
	<u>Ours</u>	0.202	0.957	0.209	0.955	0.179	0.987	0.217	0.934	0.208	0.949	0.214	0.939	0.208	0.945	0.204	0.953	0.190	0.978	0.207	0.950	0.211	0.945	0.178	0.986
	MCos	0.197	0.960	0.201	0.961	0.159	0.992	0.213	0.937	0.201	0.959	0.211	0.942	0.205	0.951	0.208	0.950	0.188	0.980	0.207	0.951	0.210	0.947	0.159	0.992
	KNN	0.195	0.963	0.203	0.955	0.162	0.992	0.207	0.946	0.197	0.966	0.207	0.947	0.205	0.951	0.205	0.955	0.188	0.981	0.203	0.957	0.206	0.952	0.157	0.992
	Maha	0.194	0.964	0.199	0.964	0.163	0.991	0.210	0.942	0.199	0.962	0.207	0.948	0.200	0.960	0.203	0.957	0.187	0.984	0.204	0.956	0.208	0.951	0.158	0.991

Table 7: Comparative analysis of AURC in the black-box SCOD framework between 12 existing OOD detection methods and our method (combining the other 12 methods) for a few models trained on ImageNet. Results are sorted in descending order by average AURC.

		Avg		iNat.		NINCO		OpenIm.		Places		Species		SSB (e)		SSB (h)		Tex.	
		RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC
DenseNet-121 (ImageNet)	Maha	0.439	0.686	0.441	0.657	0.440	0.689	0.439	0.695	0.475	0.622	0.428	0.694	0.402	0.743	0.491	0.626	0.395	0.764
	ReAct	0.385	0.810	0.331	0.899	0.416	0.751	0.381	0.821	0.347	0.869	0.383	0.831	0.422	0.729	0.452	0.713	0.349	0.866
	KNN	0.366	0.825	0.358	0.831	0.376	0.804	0.355	0.846	0.381	0.796	0.367	0.819	0.335	0.874	0.447	0.699	0.311	0.928
	RMaha	0.354	0.849	0.329	0.874	0.360	0.838	0.355	0.854	0.368	0.837	0.333	0.881	0.337	0.863	0.400	0.793	0.350	0.854
	MCos	0.350	0.872	0.310	0.929	0.368	0.838	0.342	0.890	0.347	0.873	0.355	0.870	0.337	0.888	0.427	0.750	0.314	0.940
	Energy	0.348	0.875	0.307	0.928	0.365	0.840	0.332	0.897	0.334	0.909	0.356	0.867	0.339	0.883	0.413	0.784	0.341	0.894
	ODIN	0.345	0.884	0.314	0.931	0.355	0.857	0.331	0.902	0.334	0.913	0.352	0.875	0.334	0.894	0.407	0.799	0.336	0.899
	<u>Ours</u>	0.341	0.882	0.314	0.909	0.353	0.864	0.329	0.899	0.340	0.876	0.338	0.889	0.329	0.895	0.404	0.800	0.318	0.920
	KL M	0.340	0.883	0.314	0.923	0.351	0.866	0.344	0.884	0.322	0.897	0.339	0.886	0.312	0.905	0.406	0.810	0.328	0.894
	IGEOOD	0.335	0.881	0.306	0.929	0.351	0.851	0.319	0.902	0.318	0.910	0.343	0.871	0.318	0.905	0.399	0.785	0.329	0.898
	MaxL	0.334	0.880	0.307	0.928	0.345	0.851	0.317	0.901	0.319	0.914	0.342	0.870	0.322	0.892	0.400	0.789	0.323	0.898
	Doctor	0.323	0.896	0.305	0.933	0.330	0.880	0.323	0.896	0.319	0.915	0.312	0.904	0.310	0.911	0.351	0.845	0.332	0.888
	MSP	0.322	0.892	0.298	0.929	0.331	0.876	0.324	0.891	0.317	0.911	0.314	0.901	0.311	0.908	0.354	0.837	0.331	0.882
MobileNetV3-L (ImageNet)	Maha	0.440	0.676	0.419	0.691	0.431	0.684	0.427	0.709	0.466	0.617	0.436	0.675	0.409	0.728	0.527	0.575	0.405	0.728
	KNN	0.383	0.763	0.414	0.683	0.384	0.767	0.381	0.788	0.425	0.687	0.369	0.777	0.351	0.825	0.402	0.740	0.340	0.837
	MCos	0.338	0.839	0.321	0.857	0.356	0.807	0.340	0.852	0.353	0.799	0.327	0.851	0.326	0.859	0.375	0.780	0.301	0.905
	<u>Ours</u>	0.327	0.874	0.304	0.893	0.340	0.860	0.316	0.898	0.332	0.859	0.319	0.882	0.318	0.898	0.380	0.796	0.308	0.906
	RMaha	0.322	0.865	0.302	0.906	0.333	0.846	0.318	0.883	0.318	0.864	0.303	0.896	0.318	0.860	0.374	0.788	0.308	0.879
	KL M	0.315	0.903	0.295	0.930	0.334	0.877	0.315	0.907	0.296	0.919	0.302	0.917	0.305	0.908	0.378	0.836	0.299	0.928
	ODIN	0.309	0.909	0.293	0.934	0.329	0.877	0.302	0.920	0.297	0.938	0.304	0.913	0.308	0.908	0.337	0.860	0.301	0.924
	Energy	0.302	0.911	0.278	0.947	0.328	0.873	0.290	0.929	0.292	0.938	0.298	0.913	0.298	0.916	0.346	0.834	0.287	0.938
	Doctor	0.302	0.909	0.292	0.927	0.313	0.891	0.299	0.913	0.298	0.924	0.293	0.917	0.302	0.912	0.327	0.866	0.292	0.926
	IGEOOD	0.301	0.909	0.289	0.935	0.324	0.873	0.289	0.925	0.285	0.935	0.300	0.909	0.295	0.917	0.337	0.845	0.290	0.930
	ReAct	0.300	0.903	0.276	0.944	0.318	0.870	0.286	0.924	0.289	0.924	0.299	0.906	0.293	0.912	0.357	0.802	0.278	0.940
	MSP	0.298	0.902	0.283	0.927	0.312	0.884	0.296	0.903	0.293	0.916	0.290	0.909	0.298	0.904	0.323	0.858	0.289	0.916
	MaxL	0.292	0.918	0.275	0.950	0.313	0.885	0.281	0.935	0.281	0.945	0.288	0.922	0.287	0.924	0.335	0.841	0.275	0.945
MobileNetV3-S (ImageNet)	KNN	0.473	0.694	0.507	0.623	0.466	0.701	0.472	0.700	0.515	0.622	0.461	0.705	0.438	0.758	0.493	0.678	0.429	0.761
	Maha	0.452	0.737	0.448	0.731	0.452	0.732	0.449	0.748	0.466	0.700	0.447	0.737	0.423	0.791	0.518	0.647	0.410	0.808
	MCos	0.419	0.794	0.402	0.822	0.428	0.773	0.418	0.802	0.448	0.734	0.406	0.812	0.410	0.809	0.462	0.745	0.382	0.858
	ReAct	0.412	0.808	0.387	0.852	0.418	0.790	0.405	0.818	0.394	0.841	0.415	0.807	0.398	0.828	0.491	0.691	0.389	0.841
	ODIN	0.405	0.862	0.385	0.884	0.413	0.846	0.396	0.866	0.399	0.874	0.401	0.868	0.408	0.862	0.437	0.827	0.401	0.871
	<u>Ours</u>	0.404	0.835	0.392	0.834	0.409	0.825	0.398	0.848	0.413	0.814	0.396	0.846	0.389	0.864	0.454	0.775	0.378	0.877
	IGEOOD	0.400	0.853	0.388	0.869	0.406	0.838	0.387	0.861	0.394	0.863	0.403	0.851	0.400	0.855	0.432	0.817	0.393	0.870
	KL M	0.398	0.863	0.392	0.881	0.394	0.857	0.405	0.861	0.391	0.867	0.389	0.872	0.386	0.869	0.450	0.810	0.374	0.888
	RMaha	0.390	0.831	0.381	0.861	0.395	0.818	0.392	0.830	0.390	0.826	0.376	0.852	0.377	0.841	0.433	0.773	0.376	0.848
	Energy	0.388	0.854	0.364	0.890	0.398	0.833	0.377	0.863	0.377	0.877	0.382	0.856	0.379	0.866	0.459	0.761	0.367	0.884
	Doctor	0.381	0.871	0.375	0.887	0.386	0.859	0.373	0.876	0.374	0.885	0.373	0.877	0.379	0.876	0.413	0.828	0.378	0.880
	MSP	0.377	0.873	0.362	0.892	0.382	0.862	0.369	0.876	0.369	0.887	0.372	0.876	0.375	0.879	0.411	0.831	0.372	0.883
	MaxL	0.376	0.884	0.365	0.903	0.381	0.872	0.367	0.890	0.365	0.902	0.371	0.888	0.366	0.896	0.435	0.810	0.358	0.906
ResNet-101 (ImageNet)	Maha	0.378	0.782	0.343	0.835	0.403	0.736	0.352	0.823	0.379	0.764	0.384	0.773	0.370	0.791	0.472	0.644	0.318	0.893
	KL M	0.326	0.900	0.313	0.919	0.339	0.881	0.321	0.906	0.301	0.924	0.343	0.885	0.294	0.928	0.380	0.844	0.311	0.914
	Energy	0.325	0.890	0.302	0.912	0.343	0.861	0.313	0.905	0.313	0.919	0.334	0.876	0.320	0.903	0.363	0.830	0.313	0.913
	KNN	0.323	0.877	0.310	0.893	0.334	0.853	0.310	0.899	0.313	0.892	0.329	0.868	0.307	0.903	0.399	0.757	0.282	0.948
	ODIN	0.320	0.890	0.308	0.914	0.334	0.863	0.308	0.902	0.308	0.917	0.328	0.874	0.314	0.903	0.356	0.833	0.307	0.912
	MCos	0.319	0.892	0.288	0.940	0.335	0.860	0.307	0.912	0.307	0.909	0.324	0.886	0.310	0.905	0.391	0.780	0.286	0.949
	ReAct	0.319	0.889	0.288	0.938	0.339	0.852	0.307	0.905	0.295	0.932	0.324	0.885	0.330	0.864	0.363	0.817	0.301	0.917
	<u>Ours</u>	0.316	0.910	0.284	0.944	0.333	0.885	0.305	0.930	0.307	0.927	0.320	0.902	0.312	0.921	0.371	0.832	0.299	0.942
	IGEOOD	0.315	0.902	0.302	0.917	0.323	0.885	0.302	0.916	0.301	0.928	0.325	0.885	0.303	0.924	0.356	0.842	0.307	0.917
	MaxL	0.314	0.893	0.302	0.909	0.327	0.869	0.301	0.906	0.300	0.923	0.325	0.877	0.306	0.908	0.351	0.834	0.301	0.915
	RMaha	0.312	0.893	0.295	0.925	0.318	0.881	0.307	0.905	0.314	0.886	0.306	0.906	0.309	0.890	0.343	0.850	0.302	0.903
	MSP	0.311	0.894	0.295	0.916	0.320	0.879	0.309	0.895	0.304	0.915	0.310	0.888	0.305	0.910	0.333	0.854	0.315	0.894
	Doctor	0.310	0.900	0.302	0.914	0.319	0.884	0.305	0.905	0.303	0.920	0.308	0.895	0.303	0.916	0.330	0.862	0.314	0.900
ResNet-34 (ImageNet)	Maha	0.433	0.719	0.411	0.741	0.437	0.713	0.425	0.740	0.461	0.667	0.425	0.727	0.407	0.760	0.506	0.628	0.395	0.779
	KNN	0.368	0.828	0.364	0.834	0.377	0.810	0.365	0.835	0.365	0.825	0.372	0.823	0.343	0.865	0.446	0.709	0.315	0.919
	ODIN	0.353	0.880	0.331	0.913	0.367	0.854	0.344	0.891	0.336	0.911	0.355	0.876	0.348	0.888	0.406	0.807	0.341	0.899
	ReAct	0.353	0.872	0.315	0.927	0.379	0.830	0.340	0.892	0.327	0.919	0.358	0.876	0.361	0.848	0.414	0.777	0.331	0.908
	KL M	0.352	0.887	0.329	0.920	0.354	0.879	0.355	0.885	0.341	0.899	0.366	0.883	0.324	0.908	0.409	0.824	0.342	0.895
	<u>Ours</u>	0.352	0.886	0.317	0.919	0.365	0.869	0.343	0.901	0.347	0.889	0.352	0.885	0.342	0.903	0.415	0.801	0.332	0.919
	MCos	0.350	0.873	0.317	0.922	0.364	0.844	0.347	0.883	0.343	0.880	0.341	0.879	0.341	0.880	0.433	0.756	0.311	0.939
	RMaha	0.346	0.862	0.323	0.902	0.354	0.851	0.343	0.870	0.350	0.856	0.327	0.893	0.338	0.866	0.391	0.801	0.346	0.853
	Energy	0.345	0.876	0.322	0.909	0.360	0.846	0.334	0.889	0.325	0.913	0.349	0.869	0.341	0.883	0.400	0.801	0.332	0.901
	IGEOOD	0.340	0.883	0.325	0.916	0.349	0.861	0.330	0.893	0.323	0.911	0.350	0.870	0.324	0.905	0.392	0.803	0.327	0.904
	MaxL	0.338	0.882	0.319	0.916	0.350	0.857	0.327	0.893	0.318	0.915	0.341	0.879	0.330	0.890	0.395	0.803	0.322	0.906
	Doctor	0.334	0.890	0.316	0.922	0.338	0.880	0.337	0.884	0.328	0.907	0.327	0.896	0.324	0.906	0.360	0.844	0.341	0.884
	MSP	0.332	0.885	0.309	0.922	0.337	0.875	0.332	0.880	0.325	0.902	0.328	0.889	0.324	0.898	0.360	0.835	0.338	0.879
ResNet-50 (ImageNet)	Maha	0.398	0.751	0.375	0.776	0.412	0.722	0.386	0.774	0.416	0.703	0.393	0.756	0.385	0.766	0.481	0.645	0.334	0.870
	KL M	0.338	0.888	0.319	0.918	0.348	0.869	0.337	0.898	0.338	0.897	0.328	0.897	0.307	0.914	0.395	0.819	0.334	0.891
	Energy	0.338	0.878	0.305	0.918	0.356	0.846	0.319	0.902	0.328	0.908	0.342	0.870	0.325	0.895	0.399	0.790	0.333	0.895
	KNN	0.337	0.858	0.331	0.871	0.349	0.832	0.324	0.881	0.331	0.861	0.336	0.856	0.312	0.891	0.421	0.729	0.289	0.941
	ODIN	0.333	0.886	0.311	0.923	0.346	0.856	0.313	0.909	0.324	0.913	0.338	0.882	0.317	0.905	0.388	0.805	0.331	0.896
	ReAct	0.330	0.876	0.295	0.933	0.353	0.834	0.312	0.899	0.304	0.926	0.337	0.874	0.335	0.853	0.386	0.790	0.317	0.897
	MaxL	0.329	0.882	0.308	0.918	0.341	0.855	0.310	0.904	0.316	0.912	0.333	0.877	0.312	0.903	0.388	0.796	0.325	0.896
	MCos	0.329	0.885	0.297	0.935	0.345	0.852	0.317	0.907	0.321	0.895	0.328	0.886	0.318	0.898	0.407	0.767	0.296	0.943
	RMaha	0.327	0.874	0.309	0.903	0.334	0.864	0.323	0.885	0.330	0.866	0.315	0.896	0.317	0.877	0.365	0.829	0.324	0.872
	IGEOOD	0.323	0.894	0.304	0.925	0.333	0.874	0.305	0.918	0.311	0.918	0.326	0.886	0.303	0.923	0.380	0.811	0.325	0.900
	<u>Ours</u>	0.320	0.900	0.293	0.934	0.335	0.877	0.306	0.920	0.314	0.904	0.316	0.903	0.309	0.916	0.388	0.813	0.302	0.932
	Doctor	0.319	0.890	0.303	0.923	0.329	0.868	0.310	0.897	0.317	0.903	0.305	0.897	0.303	0.909	0.342	0.842	0.340	0.878
	MSP	0.318	0.894	0.294	0.927	0.329	0.874	0.310	0.901	0.317	0.908	0.308	0.901	0.304	0.913	0.346	0.843	0.337	0.882

Table 8: Comparative analysis of AURC in the black-box SCOD framework between 12 existing OOD detection methods and our method (combining the other 12 methods) for vision transformers trained on ImageNet. Results are sorted in descending order by average AURC.

		AVG		INAT.		NINCO		OPENIM.		PLACES		SPECIES		SSB (E)		SSB (H)		TEX.	
		RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC	RC	ROC
VIT-B/16 (IMAGENET)	KL M	0.271	0.920	0.236	0.965	0.284	0.902	0.257	0.939	0.266	0.925	0.278	0.909	0.260	0.932	0.324	0.860	0.265	0.930
	KNN	0.263	0.919	0.245	0.955	0.275	0.895	0.251	0.941	0.263	0.920	0.273	0.904	0.251	0.941	0.299	0.854	0.251	0.942
	OURS	0.262	0.936	0.233	0.969	0.277	0.914	0.252	0.954	0.260	0.940	0.267	0.929	0.257	0.948	0.292	0.888	0.261	0.942
	REACT	0.257	0.926	0.237	0.962	0.269	0.902	0.245	0.948	0.251	0.935	0.271	0.910	0.245	0.946	0.284	0.876	0.256	0.932
	ENERGY	0.255	0.940	0.234	0.967	0.268	0.920	0.245	0.956	0.251	0.947	0.273	0.919	0.245	0.956	0.273	0.908	0.251	0.947
	DOCTOR	0.255	0.935	0.235	0.969	0.267	0.916	0.246	0.949	0.255	0.936	0.259	0.929	0.250	0.942	0.270	0.903	0.255	0.935
	MSP	0.254	0.936	0.234	0.968	0.266	0.916	0.245	0.950	0.254	0.938	0.258	0.928	0.251	0.944	0.270	0.905	0.256	0.937
	MCOS	0.254	0.935	0.238	0.964	0.265	0.913	0.242	0.955	0.248	0.943	0.266	0.918	0.244	0.953	0.281	0.887	0.246	0.950
	MAXL	0.253	0.943	0.240	0.967	0.261	0.927	0.244	0.956	0.250	0.948	0.270	0.922	0.243	0.957	0.268	0.914	0.247	0.951
	ODIN	0.250	0.944	0.236	0.968	0.261	0.925	0.240	0.958	0.247	0.949	0.265	0.924	0.240	0.958	0.263	0.916	0.245	0.951
	RMAHA	0.246	0.939	0.228	0.969	0.253	0.926	0.236	0.957	0.246	0.938	0.247	0.941	0.242	0.943	0.272	0.893	0.244	0.941
	IGEOOD	0.245	0.943	0.236	0.970	0.258	0.920	0.233	0.960	0.240	0.950	0.258	0.928	0.235	0.958	0.261	0.911	0.241	0.949
	MAHA	0.243	0.944	0.219	0.970	0.251	0.933	0.235	0.962	0.244	0.945	0.248	0.942	0.240	0.953	0.271	0.897	0.241	0.953
VIT-L/16 (IMAGENET)	KL M	0.256	0.934	0.225	0.971	0.265	0.920	0.246	0.947	0.252	0.937	0.264	0.920	0.248	0.943	0.296	0.886	0.247	0.944
	OURS	0.246	0.943	0.223	0.975	0.257	0.925	0.239	0.956	0.246	0.941	0.247	0.938	0.242	0.951	0.268	0.907	0.243	0.950
	KNN	0.245	0.936	0.233	0.963	0.257	0.911	0.233	0.955	0.239	0.943	0.254	0.923	0.233	0.955	0.275	0.884	0.235	0.955
	MCOS	0.241	0.947	0.226	0.969	0.253	0.924	0.234	0.960	0.236	0.955	0.250	0.932	0.235	0.959	0.261	0.914	0.236	0.959
	MSP	0.240	0.946	0.224	0.975	0.246	0.935	0.233	0.957	0.240	0.946	0.246	0.936	0.237	0.952	0.257	0.919	0.240	0.949
	DOCTOR	0.240	0.945	0.224	0.976	0.249	0.930	0.233	0.954	0.239	0.945	0.245	0.936	0.235	0.951	0.255	0.916	0.237	0.948
	ENERGY	0.239	0.953	0.219	0.974	0.250	0.936	0.232	0.967	0.236	0.959	0.252	0.935	0.233	0.964	0.258	0.922	0.234	0.963
	REACT	0.236	0.947	0.218	0.972	0.241	0.939	0.229	0.959	0.254	0.916	0.234	0.954	0.233	0.952	0.245	0.931	0.235	0.949
	IGEOOD	0.236	0.950	0.225	0.974	0.247	0.930	0.225	0.965	0.232	0.954	0.245	0.935	0.228	0.961	0.253	0.919	0.230	0.958
	MAXL	0.235	0.953	0.229	0.975	0.241	0.939	0.227	0.966	0.231	0.958	0.248	0.937	0.228	0.964	0.251	0.924	0.227	0.964
	ODIN	0.234	0.956	0.226	0.975	0.244	0.941	0.226	0.969	0.230	0.962	0.245	0.940	0.228	0.967	0.248	0.931	0.227	0.967
	RMAHA	0.231	0.954	0.218	0.975	0.235	0.944	0.223	0.967	0.229	0.954	0.232	0.953	0.228	0.957	0.249	0.928	0.232	0.951
	MAHA	0.229	0.954	0.209	0.974	0.237	0.940	0.222	0.967	0.228	0.957	0.233	0.951	0.226	0.959	0.251	0.924	0.226	0.961
VIT-S/16 (IMAGENET)	KL M	0.295	0.903	0.257	0.956	0.305	0.886	0.286	0.917	0.290	0.903	0.301	0.899	0.277	0.917	0.355	0.836	0.289	0.911
	KNN	0.291	0.901	0.272	0.939	0.300	0.879	0.280	0.922	0.296	0.888	0.292	0.902	0.281	0.920	0.332	0.829	0.277	0.929
	OURS	0.283	0.917	0.253	0.962	0.298	0.893	0.274	0.934	0.279	0.919	0.286	0.912	0.275	0.931	0.320	0.858	0.276	0.928
	ENERGY	0.280	0.926	0.253	0.961	0.296	0.899	0.270	0.944	0.270	0.941	0.292	0.908	0.275	0.938	0.307	0.880	0.274	0.936
	DOCTOR	0.278	0.917	0.253	0.959	0.288	0.901	0.273	0.926	0.277	0.916	0.280	0.913	0.273	0.931	0.302	0.872	0.277	0.919
	MSP	0.278	0.915	0.252	0.958	0.288	0.898	0.275	0.924	0.276	0.914	0.279	0.911	0.273	0.928	0.301	0.870	0.276	0.916
	REACT	0.276	0.915	0.259	0.946	0.287	0.893	0.268	0.933	0.267	0.927	0.298	0.880	0.260	0.942	0.303	0.866	0.267	0.930
	MCOS	0.275	0.926	0.254	0.960	0.290	0.899	0.266	0.944	0.270	0.932	0.278	0.921	0.269	0.938	0.310	0.867	0.264	0.947
	RMAHA	0.275	0.915	0.252	0.959	0.285	0.898	0.265	0.935	0.274	0.914	0.268	0.929	0.272	0.917	0.313	0.851	0.269	0.920
	MAXL	0.273	0.935	0.256	0.964	0.284	0.913	0.266	0.949	0.263	0.948	0.286	0.918	0.267	0.947	0.296	0.893	0.265	0.946
	ODIN	0.272	0.934	0.259	0.962	0.283	0.915	0.266	0.948	0.262	0.948	0.282	0.918	0.266	0.947	0.294	0.895	0.267	0.943
	IGEOOD	0.269	0.922	0.257	0.963	0.283	0.894	0.258	0.943	0.258	0.935	0.281	0.905	0.257	0.942	0.300	0.864	0.261	0.934
	MAHA	0.269	0.922	0.245	0.962	0.278	0.906	0.256	0.945	0.269	0.918	0.266	0.930	0.263	0.931	0.308	0.855	0.263	0.930
VIT-T/16 (IMAGENET)	KNN	0.357	0.831	0.337	0.866	0.369	0.810	0.340	0.862	0.369	0.802	0.355	0.832	0.336	0.864	0.426	0.731	0.325	0.884
	KL M	0.345	0.882	0.302	0.937	0.361	0.860	0.339	0.899	0.326	0.894	0.351	0.877	0.332	0.888	0.425	0.796	0.321	0.905
	MAHA	0.342	0.862	0.303	0.923	0.350	0.851	0.326	0.894	0.357	0.828	0.332	0.880	0.330	0.881	0.400	0.785	0.343	0.853
	OURS	0.335	0.897	0.296	0.942	0.352	0.872	0.321	0.920	0.331	0.901	0.337	0.895	0.326	0.912	0.395	0.818	0.322	0.916
	RMAHA	0.332	0.876	0.300	0.937	0.345	0.856	0.317	0.904	0.334	0.866	0.318	0.898	0.326	0.880	0.397	0.784	0.321	0.884
	MCOS	0.331	0.882	0.301	0.924	0.350	0.851	0.317	0.907	0.331	0.878	0.330	0.880	0.318	0.901	0.393	0.794	0.308	0.921
	IGEOOD	0.329	0.891	0.308	0.932	0.353	0.852	0.311	0.919	0.315	0.910	0.345	0.866	0.307	0.923	0.395	0.792	0.302	0.931
	ODIN	0.326	0.904	0.303	0.944	0.347	0.867	0.311	0.928	0.314	0.925	0.339	0.887	0.312	0.922	0.375	0.825	0.306	0.935
	REACT	0.325	0.896	0.288	0.952	0.344	0.862	0.305	0.927	0.311	0.919	0.336	0.879	0.317	0.904	0.388	0.800	0.307	0.923
	MSP	0.322	0.887	0.299	0.930	0.337	0.863	0.316	0.895	0.319	0.893	0.321	0.884	0.316	0.894	0.354	0.832	0.310	0.902
	ENERGY	0.321	0.906	0.293	0.947	0.345	0.867	0.304	0.933	0.307	0.929	0.336	0.888	0.307	0.925	0.377	0.822	0.300	0.939
	DOCTOR	0.321	0.899	0.303	0.933	0.334	0.881	0.316	0.907	0.319	0.906	0.319	0.898	0.315	0.908	0.353	0.846	0.308	0.915
	MAXL	0.314	0.907	0.294	0.946	0.334	0.872	0.299	0.928	0.300	0.928	0.328	0.891	0.299	0.926	0.367	0.824	0.292	0.939