Out-of-Distribution Detection with Overlap Index

Anonymous authors
Paper under double-blind review

Abstract

Out-of-distribution (OOD) detection is crucial for the deployment of machine learning models in the open world. While existing OOD detectors are effective in identifying OOD samples that deviate significantly from in-distribution (ID) data, they often come with trade-offs. For instance, deep OOD detectors usually suffer from high computational costs, require tuning hyperparameters, and have limited interpretability, whereas traditional OOD detectors may have a low accuracy on large high-dimensional datasets. To address these limitations, we propose a novel effective OOD detection approach that employs an overlap index (OI)-based confidence score function to evaluate the likelihood of a given input belonging to the same distribution as the available ID samples. The proposed OI-based confidence score function is non-parametric, lightweight, and easy to interpret, hence providing strong flexibility and generality. Extensive empirical evaluations indicate that our OI-based OOD detector is competitive with state-of-the-art OOD detectors in terms of detection accuracy on a wide range of datasets while requiring less computation and memory costs. Lastly, we show that the proposed OI-based confidence score function inherits nice properties from OI (e.g., insensitivity to small distributional variations and robustness against Huber ϵ-contamination) and is a versatile tool for estimating OI and model accuracy in specific contexts.

1 Introduction

Machine learning models often struggle with out-of-distribution (OOD) samples that originate from distributions not seen during training. The OOD detection (Macêdo et al., 2021; Zhao et al., 2023) is a critical task to address the above issue and has been utilized in many real-world applications (Zhao et al., 2023; Wang et al., 2022a). Given in-distribution (ID) samples originating from the training distribution, the OOD detector employs a confidence scorer to quantify the distance between a given input and the available ID samples to filter out OOD inputs, thereby enhancing the machine learning model's reliability in unseen environments.

Traditional OOD detectors like one-class SVM (Schölkopf et al., 2001) are computationally and memory-efficient but are relatively less effective on large high-dimensional datasets. While deep learning approaches such as Deep SVDD (Ruff et al., 2018) are capable of handling large high-dimensional datasets, they come at the cost of substantial computational and memory overhead due to their large number of parameters. Gaussian-based OOD detectors, such as (Morteza & Li, 2022), require calculating the inverse of the covariance matrix which can be numerically unstable. To address these limitations, this paper proposes a novel OOD detector that is non-parametric, lightweight, and effective on large high-dimensional datasets.

The proposed OOD detector employs an overlap index (OI)-based confidence score function to evaluate the likelihood of a given input belonging to the same distribution as the available ID samples. OI is a widely utilized metric that quantifies the area of intersection between two probability density functions. It is closely related to the total variation distance (TVD), where TVD is one minus OI (Pastore & Calcagnì, 2019). Both OI and TVD find applications in various domains (Zhang et al., 2021b; Kapralov et al., 2016). Unlike Kullback–Leibler (KL) and Jensen–Shannon (JS) divergences, OI does not require the supports of the distributions to be identical and is insensitive to small distributional variations (Larry Wasserman, a). While the Wasserstein distance requires distance inference (Yatracos, 2022) and lacks robustness to Huber ϵ-contamination outliers (Huber, 1992) with unbounded metrics, OI does not suffer from these limitations. Despite the applications and advantages of OI, its utilization for OOD detection is seldom leveraged in

existing literature. This paper endeavors to explore the potential applications of the OI-based confidence score function in OOD detection, showcasing its effectiveness.

Our OI-based confidence score function is converted from a novel upper bound of OI derived in this paper. The derived bound consists of two computationally efficient terms: the distance between the means of the clusters and the TVD between two distributions over subsets that meet specific conditions. The derived upper bound is quick to compute, non-parametric, insensitive to small distributional variations, and robust against Huber ϵ-contamination outliers. If one cluster contains only a given input and the other cluster contains a few clean ID samples, then this bound could serve as a confidence score function for evaluating the likelihood that the given input belongs to the same distribution as the available ID samples.

The proposed OI-based OOD detector addresses the aforementioned limitations of the existing OOD detectors. Compared to traditional OOD detectors, our method is empirically effective even on large high-dimensional datasets. Compared to deep OOD detectors, our OI-based detector avoids the use of neural networks, thereby being more economical in terms of computational and memory requirements. Despite this, it achieves accuracy levels comparable to those of state-of-the-art deep OOD detectors. This makes our approach particularly beneficial in scenarios where the available ID dataset is too small to effectively train a deep learning-based OOD detector. Another group of OOD detectors is Gaussian-based OOD detectors, such as GEM (Morteza & Li, 2022), MSP (Hendrycks & Gimpel, 2017), Mahalanobis distance (Lee et al., 2018), and energy score (Liu et al., 2020). These methods assume Gaussian distributions and apply distance-based metrics on deep features. Compared to them, our method neither makes distributional assumptions about the underlying feature space nor requires the computationally challenging and numerically unstable task of inverting covariance matrices. Zhang et al. (2023) uses an KL-based confidence scorer for OOD detection, which also assumes Gaussian distribution for the extracted features.

We show that the variant of our OI-based confidence scorer could be employed for estimating both OI and model accuracy in specific contexts. Sriperumbudur et al. (2012) estimates OI using integral probability metrics (IPM)[1], which results in a computational complexity on the order of $\mathcal{O}^*(n^\omega)$, where ω is the matrix multiplication exponent (Cohen et al., 2021). Additionally, selecting appropriate functions for IPM-based methods often presents a challenge (Larry Wasserman, b). In contrast, our method approximates OI using a class of straightforward, easily identifiable, and computationally efficient conditional functions[2]. Schmid & Schmidt (2006); Pastore & Calcagnì (2019) estimate OI using estimated probability density functions through kernel methods[3]. These methods are burdened by the curse of dimensionality (Zambom & Ronaldo, 2013), whereas our approach remains efficient in high-dimensional spaces.

Overall, the contributions of this paper include: 1) proposing a novel OOD detector that leverages an OI-based confidence score function; 2) evaluating the proposed OOD detector with various state-of-the-art methods and datasets to demonstrate its effectiveness and efficiency; 3) analyzing the mathematical properties of the OI-based confidence score function; and 4) exploring into the potential applications of the OI-based confidence score function for estimating OI and model accuracy in specific contexts.

The remaining part is organized as follows: We first conduct a literature review on OI and OOD detection in Sec. 2. Then Sec. 3 derives the novel upper bound for OI and proposes our novel OI-based OOD detector. Sec. 4 empirically evaluates the proposed novel OOD detector. Sec. 5 provides mathematical analysis and the potential applications of our approach for estimating OI and model accuracy. Sec. 6 discusses the limitations and broader impact of our approach. Sec. 7 concludes the paper.

2 Background and Related Works

2.1 OOD Detection

A comprehensive discussion of classic OOD detectors, such as one-class SVM (Schölkopf et al., 2001), decision-tree (Comité et al., 1999), and one-class nearest neighbor (Tax, 2002), is given by Khan & Madden (2014).

[1] $d_{\mathcal{F}}(P,Q) = \sup_{f\in\mathcal{F}} |\mathbb{E}X \sim P[f(X)] - \mathbb{E}Y \sim Q[f(Y)]|$, where $\mathcal{F}$ is a class of functions.

[2] The condition function $\mathbb{1}\{\cdot\}$ outputs 1 when the input satisfies the given condition and 0 otherwise.

[3] $\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}\left(\frac{x-x_i}{\beta}\right)$, where $\mathcal{K}$ is the kernel function and β is the bandwidth.

The deep OOD detectors can be divided into image-level OOD detectors and feature-level OOD detectors. The image-level OOD detectors train their models using raw inputs, whereas the feature-level OOD detectors require pretrained models to process the input. For image-level OOD detectors, Deep SVDD (Ruff et al., 2018), OCGAN (Perera et al., 2019), and GradCon (Kwon et al., 2020) do not use additional information, whereas Deep SAD (Ruff et al., 2020) uses information from anomaly samples. Bergmann et al. (2020) uses a teacher network for OOD detection. CutPaste (Li et al., 2021a) considers object detection. Schneider et al. (2022) studies different autoencoders for OOD detection. For feature-level OOD detectors, You et al. (2023) achieves a higher accuracy using a transformer instead of autoencoders to reconstruct feature maps. Golan & El-Yaniv (2018) bypasses the feature reconstruction phase using self-labeled training datasets. LOE (Qiu et al., 2022), Panda (Reiss et al., 2021), and Salehi et al. (2021) use pretrained ImageNet models to increase their accuracy. LPIPS (Zhang et al., 2018) is a feature-based similarity metric for images. ECOD (Li et al., 2022) is a non-parametric OOD detector, which calculates the confidence score using empirical cumulative distribution functions. Xu et al. (2023) proposes a deep isolation forest for OOD detection that uses neural networks to map original data into random representation ensembles. Other works aiming to improve OOD detection accuracy for classification and object detection include ViM (Wang et al., 2022b), GradNorm (Huang et al., 2021), VOS (Du et al., 2022), PatchCore (Roth et al., 2022), YolOOD (Zolfi et al., 2024). Li et al. (2023) uses masked image modeling for OOD detection.

2.2 Statistical Divergences and Estimating OI

Popular statistical divergences to measure distribution similarities contain OI, TVD, KL divergence, and Wasserstein distance. KL divergence is often utilized in classification (Malinin & Gales, 2019) and reinforcement learning (Even-Dar et al., 2006), whereas the Wasserstein distance is preferred in optimal transport (Kolouri et al., 2017; Raghvendra et al., 2024; Phatak et al., 2023; Lahn et al., 2023) and image processing (Cuturi & Doucet, 2014). Note that TVD is less than or equal to the square root of KL divergence. Therefore, many results from KL divergence can transfer to TVD and thus OI, or vice versa. Estimating OI and TVD with unknown distributions using only finite samples is challenging. The kernel-based estimations (Schmid & Schmidt, 2006; Pastore & Calcagnì, 2019) are sensitive to the choice of bandwidth, have a boundary effect, and suffer from the curse of dimensionality. We found that their methods are slow in high-dimensional space. The Cohen's d measure (Inman & Bradley Jr, 1989) is efficient but assumes that the distributions A, B are Gaussian with the same standard deviation σ and approximates $OI \approx 2\Phi(-\frac{|\mu_A - \mu_B|}{2\sigma})$, where μ_A and μ_B are samples' means and $\Phi(\cdot)$ is the standard normal distribution function. If the assumption does not hold, the method may not be accurate. Sriperumbudur et al. (2012) provides an estimator for TVD, which requires solving time-consuming linear programs and is not consistently accurate, as noted by the authors. Our OI-based confidence score function shows a potential for estimating OI in a time-efficient fashion without suffering from the curse of dimensionality.

3 The OI-Based OOD Detector

3.1 Problem Formulation and Goals

Given $\mathbb{R}^n$ space and m samples $\{x_i\}_{i=1}^m$ that lie in an unknown distribution, we would like to build a binary detector $\Psi : \mathbb{R}^n \to \{\pm 1\}$ such that for any new input x, $\Psi(x)$ outputs 1 when x is from the same unknown probability distribution, and outputs -1, otherwise. The detector Ψ should be effective even when n is large or m is small. Additionally, the computation and memory costs of Ψ should be small.

3.2 Motivation: A Novel Upper Bound for OI

Our methodology is inspired by a newly derived upper bound for the OI between bounded distributions. We consider the $\mathbb{R}^n$ space and continuous random variables. We define P and Q as two probability distributions in $\mathbb{R}^n$ with f_P and f_Q being their probability density functions. The derivation of this novel upper bound necessitates the introduction of the following definitions.

Definition 3.1. The OI, η, is a function of two distributions whose samples belong to the $\mathbb{R}^n$ space and outputs the overlap index value of these two distributions, which is a real number between 0 and 1. η is

defined as (Pastore & Calcagnì, 2019):

$$\eta(P,Q) = \int_{\mathbb{R}^n} \min[f_P(x), f_Q(x)]dx. \tag{1}$$

Definition 3.2. Given an $A \subset \mathbb{R}^n$, δ_A is a function of two distributions whose samples belong to the $\mathbb{R}^n$ space and outputs a real number between 0 and 1 with the following definition:

$$\delta_A(P,Q) = \frac{1}{2}\int_A |f_P(x) - f_Q(x)|dx. \tag{2}$$

The standard TVD is δ_A with $A = \mathbb{R}^n$ or one minus OI (Pastore & Calcagnì, 2019). The following theorem shows a novel upper bound of OI between bounded distributions.

Theorem 3.3. *Without loss of generality, assume D^+ and D^- are two probability distributions on a bounded domain $B \subset \mathbb{R}^n$ with defined norm[4] $||\cdot||$ (i.e., $\sup_{x\in B}||x|| < +\infty$), then $\forall\ A \subset B$ with $A^c = B \setminus A$, we have*

$$\eta \leq 1 - \frac{1}{2r_{A^c}}||\mu_{D^+} - \mu_{D^-}|| - \frac{r_{A^c} - r_A}{r_{A^c}}\delta_A \tag{3}$$

where $r_A = \sup_{x\in A}||x||$ and $r_{A^c} = \sup_{x\in A^c}||x||$, μ_{D^+} and μ_{D^-} are the means of D^+ and D^-, and δ_A is TVD on set A as defined in **Definition** *3.2. Moreover, let $r_B = \sup_{x\in B}||x||$, then*

$$\eta \leq 1 - \frac{1}{2r_B}||\mu_{D^+} - \mu_{D^-}|| - \frac{r_B - r_A}{r_B}\delta_A. \tag{4}$$

Since (4) holds for any A, a tighter bound can be written as

$$\eta \leq 1 - \frac{1}{2r_B}||\mu_{D^+} - \mu_{D^-}|| - \max_A \frac{r_B - r_A}{r_B}\delta_A. \tag{5}$$

The theorem is framed to accommodate probability distributions with bounded domains. In practical applications, data may originate from unbounded distributions. Nevertheless, the corresponding finite-sample datasets are inherently bounded. As such, it is possible to approximate the underlying unbounded distributions with bounded ones and then apply our theorem. We forego further exploration of cases involving unbounded distributions at this time, as the current theorem sufficiently informs the intuition behind our proposed OOD detector. To calculate δ_A and $||\mu_{D^+} - \mu_{D^-}||$ with finite samples, we have the following corollary.

Corollary 3.4. *Given D^+, D^-, B, and $||\cdot||$ used in* **Theorem** *3.3, let $A(g) = \{x \mid g(x) = 1, x \in B\}$ with any condition function $g : B \to \{0, 1\}$. An upper bound for $\eta(D^+, D^-)$ can be obtained:*

$$\eta \leq \overline{\eta} = 1 - \frac{1}{2r_B}||\mu_{D^+} - \mu_{D^-}|| - \max_g \frac{r_B - r_{A(g)}}{2r_B}\left|\mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g]\right|.$$

If we use only a single input to calculate μ_{D^+} and $\mathbb{E}_{D^+}[g]$ and use a few samples to calculate μ_{D^-} and $\mathbb{E}_{D^-}[g]$, then $\overline{\eta}$ can be considered as a confidence score function for evaluating the likelihood that the given input in D^+ belongs to the same distribution as samples in D^-. Before providing further visualization and ablation study to demonstrate the efficacy of using such a confidence score function, we present the implementation algorithm to calculate equation 6.

3.3 Empirical Calculation of the Novel Upper Bound

The middle term in the RHS of equation 6 is the distance of clusterings' means, and the last term is a function of IPM with a class of condition functions. To approximate the last term, we draw samples from D^+ and D^- and then average their g values to estimate $\mathbb{E}_{D^+}[g]$ and $\mathbb{E}_{D^-}[g]$. Alg. 1 shows the computation of $\overline{\eta}$ with given k condition functions $\{g_j\}_{j=1}^k$ and finite sample, $\{x_i^+\}_{i=1}^d \sim D^+$ and $\{x_i^-\}_{i=1}^m \sim D^-$.

[4]This paper considers the L_2 norm. However, the analysis can be carried out using other norms.

Algorithm 1 ComputeBound($\{x_i^+\}_{i=1}^d$, $\{x_i^-\}_{i=1}^m$, $\{g_j\}_{j=1}^k$)

$B \leftarrow \{x_1^+, ..., x_d^+, x_1^-, ..., x_m^-\}$
$r_B \leftarrow \max_{x \in B} ||x||$
$\Delta_\mu \leftarrow \left|\left| \frac{1}{d}\sum_{i=1}^d x_i^+ - \frac{1}{m}\sum_{i=1}^m x_i^- \right|\right|$
FOR $j = 1 \rightarrow k$
  $A(g_j) \leftarrow \{x \mid g_j(x) = 1, x \in B\}$
  $r_A \leftarrow \max_{x \in A} ||x||$
  $s_j \leftarrow (r_B - r_A) \left| \frac{1}{d}\sum_{i=1}^d g_j(x_i^+) - \frac{1}{m}\sum_{i=1}^m g_j(x_i^-) \right|$
Return: $1 - \frac{1}{2r_B}\Delta_\mu - \frac{1}{2r_B}\max_j s_j$

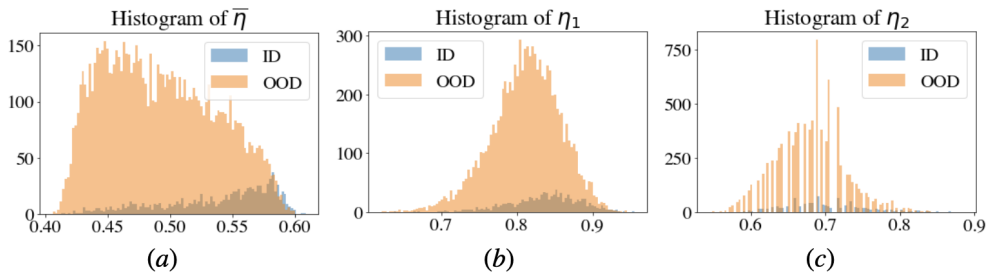

Figure 1: Histograms of confidence scores using $\overline{\eta}$, η_1, and η_2 with plane as the ID class and the other nine classes as the OOD class in CIFAR-10.

Choice of g: The choice of condition functions is not unique. The literature faces similar issues when choosing functions for IPMs (Larry Wasserman, b; Sriperumbudur et al., 2012). The overall guideline is to choose functions that satisfy convenient regularity conditions. The chosen functions should also be computation-efficient and memory-efficient. Considering all the aspects, $g_j(x) = \mathbb{1}\{r_{j-1} \leq ||x|| \leq r_j\}$ is selected with $r_j = jr_B/k$, which are nicely regularized, computation-efficient, and empirically effective. However, other condition functions are worth exploring.

The confidence score function using Alg. 1 is defined as $f(x) = \mathit{ComputeBound}(\{x\}, \{x_i\}_{i=1}^m, \{g_j\}_{j=1}^k)$ (i.e., $\{x^+\}^d = \{x\}$ with $d = 1$), which measures the maximum similarity between x and $\{x_i\}_{i=1}^m$. If $f(x) \geq T_0$, then x is considered as an ID sample. Otherwise, x is considered as OOD. T_0 is pre-defined by users.

3.4 Computation and Space Complexities

We can pre-compute and store $\frac{1}{m}\sum_{i=1}^m x_i$ and $\frac{1}{m}\sum_{i=1}^m g_j(x_i)$ in Alg. 1. Therefore, the space complexity is $\mathcal{O}(k+1)$. The computation complexity is $\mathcal{O}(k+1)$ for each online input x since it needs to calculate $||x||$ once and s_j for k times. k can be restricted to a reasonable number (e.g., ≤ 100) so that even devices without strong computation power (e.g., Arduino or Raspberry Pi) can run our OOD detector efficiently. This is an advantage compared to time-consuming deep approaches.

Visualization: Although our confidence score seems simple, we empirically find it effective. Both $\frac{1}{2r_B}||\mu_{D^+} - \mu_{D^-}||$ and $\max_g \frac{r_B - r_{A(g)}}{2r_B} |\mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g]|$ are important for our OOD detector. To validate their efficacy, we created two another OOD detectors with $\eta_1 = 1 - \frac{1}{2r_B}||\mu_{D^+} - \mu_{D^-}||$ and $\eta_2 = 1 - \max_g \frac{r_B - r_{A(g)}}{2r_B} |\mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g]|$ as the confidence score functions. Fig. 1 shows the histograms of $\overline{\eta}$, η_1, and η_2 using CIFAR-10 (Krizhevsky et al., 2009) plane images as the ID samples and the other nine class images as the OOD samples. Fig. 1(b,c) show that it is less likely to distinguish between ID and OOD samples by independently observing either $\frac{1}{2r_B}||\mu_{D^+} - \mu_{D^-}||$ or $\max_g \frac{r_B - r_{A(g)}}{2r_B} |\mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g]|$. However, the summation of $\frac{1}{2r_B}||\mu_{D^+} - \mu_{D^-}||$ and $\max_g \frac{r_B - r_{A(g)}}{2r_B} |\mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g]|$ amplifies the difference between ID and OOD samples, as shown in Fig. 1(a). Our OOD detector successfully utilizes the dependence between $\frac{1}{2r_B}||\mu_{D^+} - \mu_{D^-}||$ and $\max_g \frac{r_B - r_{A(g)}}{2r_B} |\mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g]|$ to identify OOD samples.

Table 1: Information on utilized UCI datasets.

Dataset	n	ID Class	Size	OOD Class	Size
Iris	4	Setosa	50	Vers.+Virg.	100
		Versicolour	50	Seto.+Virg.	100
		Virginica	50	Seto.+Vers.	100
Breast Cancer	9	Malignant	241	Beni.	458
		Benign	458	Mali.	241
Ecoli	7	Peripalsm	52	All Others	284
Ball-bearing	32	Ball-bearing	913	None	0

Flexibility: Our OOD detector does not require any distributional assumptions nor the need to calculate the inverse of the covariance matrix which can be numerically unstable. Besides, Alg. 1 could be applied in any space, such as the feature space generated by a pretrained model. This flexibility broadens the potential application of our approach. Additionally, Alg. 1 supports batch-wise computation since $\frac{1}{m}\sum_{i=1}^{m} x_i$ and $\frac{1}{m}\sum_{i=1}^{m} g_j(x_i)$ can be pre-computed and stored.

4 Experimental Results

4.1 Setup

Datasets: We employ UCI datasets (Dua & Graff, 2017) given in Table 1 for benchmarking against traditional OOD detectors. For comparison with state-of-the-art deep OOD detectors, we primarily use the CIFAR-10 dataset (Krizhevsky et al., 2009) in both input and feature spaces. Additionally, the CIFAR-100 dataset (Krizhevsky et al., 2009) is utilized to serve as the ID dataset for comparisons with Gaussian-based OOD detectors. OOD datasets include Textures (Cimpoi et al., 2014), SVHN (Netzer et al., 2011), LSUN-Crop (Yu et al., 2015), LSUN-Resize (Yu et al., 2015), and iSUN (Xu et al., 2015). We also extend the application of our OOD detector to backdoor detection, leveraging datasets such as MNIST (LeCun et al., 2010), GTSRB (Stallkamp et al., 2011), YouTube Face (Wolf et al., 2011), and sub-ImageNet (Deng et al., 2009). Details on these datasets are provided in Appendix .6.

Metrics: The primary metric employed for assessing the efficacy of our OOD detectors is the area under the receiver operating characteristic curve (AUROC). A higher AUROC value indicates better accuracy in differentiating between ID and OOD samples. Supplementary to AUROC, we also utilize TPR95—representing the detection accuracy for OOD samples when the detection accuracy for ID samples is fixed at 95%—and the area under the precision-recall curve (AUPR) as additional performance metrics.

Hyperparameters: We set $k = 100$ and $g_j(x)$ with the form $\mathbb{1}\{r_{j-1} \leq ||x|| \leq r_j\}$ in Alg. 1. We recommend choosing k between 50 and 200. Values of $k \geq 200$ do not significantly improve performance, while $k < 50$ results in underperformance. The selected condition functions g_j are both computationally efficient and empirically effective. The number m of available ID samples varies depending on the dataset in use. We perform ablation studies to evaluate the impact of these hyperparameters on our method's performance.

Hyperparameters of compared methods: We obtained the code for each comparison method from the authors' respective websites. The hyperparameters were carefully fine-tuned to replicate the results reported in their original papers. For instances where we applied the compared methods to new datasets, we conducted careful hyperparameter tuning to ensure their best performance.

4.2 Comparison Results

UCI Dataset: We first evaluated our OI-based detector on small UCI datasets given in Table 1. Fig. 2(a) shows that our detector outperforms all baseline methods by showing the highest average AUROC (i.e., 97.67±2.55%). Table 10 in the appendix provides the numerical numbers.

CIFAR-10 Dataset: We then use CIFAR-10 images with one class being ID and the remaining classes being OOD. The compared deep OOD detectors on raw images without using additional information from

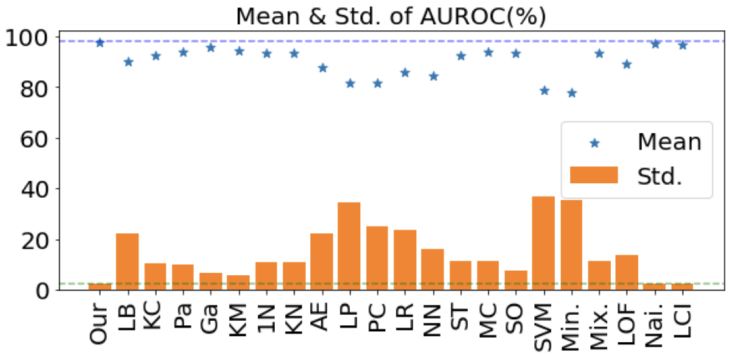

Figure 2: AUROC on UCI datasets. Horizontal lines: the mean and standard deviation of our approach.

Table 2: AUROC (%) on CIFAR-10 with each class being ID. **Boldface** shows the highest AUROC.

Class	Raw Image Space						Extra Information				
	Ours	OCGan	Deep SVDD	AnoGAN	DCAE	GradCon	Ours	LPIPS	Bergman	PANDA	ADTR
Plane	**76.7**	75.7	61.7	67.1	59.1	76.0	96.1	79.3	78.9	93.9	**96.2**
Car	**67.0**	53.1	65.9	54.7	57.4	59.8	97.5	94.6	84.9	97.1	**98**
Bird	53.4	64.0	50.8	52.9	48.9	**64.8**	94	63.1	73.4	85.4	**94.5**
Cat	58.5	**62.0**	59.1	54.5	58.4	58.6	**94.3**	73.7	74.8	85.4	91.7
Deer	73.0	72.3	60.9	65.1	54.0	**73.3**	**95.7**	72	85.1	93.6	95.1
Dog	**66.2**	62.0	65.7	60.3	62.2	60.3	**96.4**	75.2	79.3	91.2	95.6
Frog	**73.3**	72.3	67.7	58.5	51.2	68.4	97.9	83.6	89.2	94.3	**98**
Horse	58.8	57.5	**67.3**	62.5	58.6	56.7	94.9	81.8	83	93.5	**97.1**
Ship	78.0	**82.0**	75.9	75.8	76.8	78.4	97.9	81.2	86.2	95.1	**98**
Truck	**74.7**	55.4	73.1	66.5	67.3	67.8	**97.1**	86.3	84.8	95.2	96.9
Ave.	**68.0**	65.6	64.8	61.8	59.4	66.4	**96.2**	79.1	82.0	92.5	96.1

anomaly samples include DCAE (Makhzani & Frey, 2015), AnoGAN (Schlegl et al., 2017), Deep SVDD (Ruff et al., 2018), OCGAN (Perera et al., 2019), and GradCon (Kwon et al., 2020). Comparisons with more recent approaches are provided later. All the methods are given $m = 5000$ available ID samples. The result is in Table 2 "Raw Image Space". Our approach shows the highest AUROC. Although the baseline methods could reach similar accuracy, they are slower than ours on the same CPU machine. For example, the Deep SVDD requires 446.8 ms for each detection (ours is 3 ms). Fig. 3(a) shows the AUROC of our approach with m less than 5000. With only $m = 100$ ID samples, the AUROC of our approach is 67.2%. In contrast, the AUROC of Deep SVDD drops to 61.1% with $m = 100$. As for the memory cost, Deep SVDD uses LeNet (LeCun et al., 1998) with three convolutional layers. The first layer already has 2400 parameters, whereas our approach has $\mathcal{O}(k+1)$ space complexity with $k = 100$.

Ablation Study: We compare our approach with classic OOD detectors that are fit by samples' norms, including LOF (Breunig et al., 2000), OCSVM (Schölkopf et al., 2001), Isolation Forest (Liu et al., 2008), and Elliptic Envelope (Rousseeuw & Driessen, 1999). Table 3 shows that our approach is more than just calculating the statistic of the sample's norms. When using η_1 and η_2 mentioned in Sec. 3 as the confidence score functions, the AUROC of our approach becomes 58.4% for η_1 and 50.1% for η_2. These two numbers are consistent with Fig. 1 that either term cannot individually detect OOD samples. Fig. 3(b) shows the AUROC of our approach with different k. Our approach remains high AUROC when $k \geq 50$.

Table 3: AUROC of our approach using η_1 and η_2 as the confidence score functions and AUROC of other methods with samples' norms as inputs.

Method	η_1	η_2	LOF	OCSVM	Ellip. Env.	Iso. Fo
AUROC	58.4	50.1	60.5	55.2	55.2	56.7

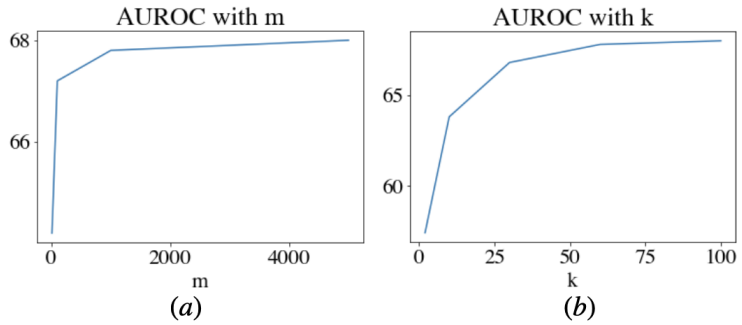

Figure 3: AUROC of our approach with (a) different numbers m of available ID samples and (b) different numbers k of condition functions.

Table 4: Execution time (ms) per sample with various data dimension.

n	10	100	500	1000	2000
Ours	3	3	3	3	3
ECOD	8	79	453	1021	2245
Deep Isolation Forest	150	150	150	150	150

Runtime Efficiency: ECOD (Li et al., 2022) is also non-parametric and does not require training parameters. The time complexity of ECOD is $\mathcal{O}(n)$ that increases with the data dimension n. We compare the execution time of our approach with ECOD by varying the data dimension n. We also report the computation time of Deep Isolation Forest (Xu et al., 2023). Table 4 shows the execution time. The computation complexity of our method and Deep Isolation Forest do not increase with the data dimension, whereas the computation complexity of ECOD increases linearly with the data dimension. The time complexity of our approach increases linearly with the number k of utilized condition functions. We empirically observed that with $k = 1000$, the execution time becomes 36 ms.

Improvement with Extra Information: Deep OOD detectors such as LPIPS (Zhang et al., 2018), Bergmann's method (Bergmann et al., 2020), Panda (Reiss et al., 2021), and ADTR (You et al., 2023) require pretrained models or additional information from anomaly samples to identify OOD samples. If our approach is also allowed to use pretrained models and information from anomaly samples, the accuracy will improve. Specifically, we use a pretrained ImageNet model from (Reiss et al., 2021) to extract CIFAR-10 features. We also assume that a contaminated dataset is available. To build the contaminated dataset, we first merge all ID and OOD samples and then randomly select 100 samples. During the testing, each input will be subtracted from the mean vector of this contaminated dataset. The results are shown in Table 2 "Extra Information". The average AUROC of our approach increases to 96.2%. Fig. 4(a) shows the improvement of

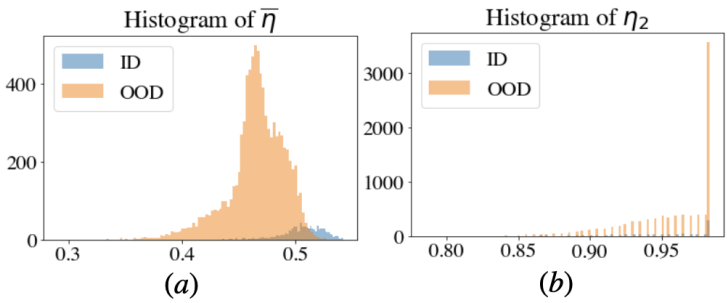

Figure 4: Histograms of confidence scores using $\overline{\eta}$ and η_2 with plane as the ID class and the other nine classes as the OOD class in CIFAR-10.

Table 5: Results for CIFAR-10 and CIFAR-100 being in-distribution datasts. **Boldface** shows the best performance, whereas underline shows the second best.

Out-of-Distribution Datasets	Method	ID Dataset: CIFAR-10 TPR95 (%)	AUROC (%)	AUPR (%)	ID Dataset: CIFAR-100 TPR95 (%)	AUROC (%)	AUPR (%)
Texture	Ours	64.20	92.80	92.33	42.50	85.79	85.27
	MSP	40.75	88.31	97.08	16.71	73.58	93.02
	Mahalanobis	62.38	94.46	98.75	57.62	90.14	97.62
	Energy Score	47.47	85.47	95.58	20.38	76.46	93.68
	GEM	**72.61**	**94.59**	**98.79**	57.40	**90.17**	97.63
	VOS	53.55	86.81	95.72	**59.08**	88.05	**97.99**
SVHN	Ours	**94.10**	**98.56**	**99.41**	**93.75**	**98.36**	**99.36**
	MSP	52.41	92.11	98.32	15.66	71.37	92.89
	Mahalanobis	79.34	95.72	99.04	51.35	89.25	97.52
	Energy Score	64.20	91.05	97.66	14.59	74.10	93.65
	GEM	79	95.65	99.01	51.51	89.40	97.57
	VOS	75.41	95.26	98.99	51.66	87.13	96.85
LSUN-Crop	Ours	83.63	96.60	96.61	57.76	89.95	90.03
	MSP	69.07	95.64	99.13	33.44	83.71	96.32
	Mahalanobis	30.06	86.15	97.05	1.53	58.48	89.73
	Energy Score	91.89	98.40	**99.67**	64.01	93.41	98.59
	GEM	30.20	86.09	97.03	1.70	58.42	89.70
	VOS	**92.45**	**98.44**	**99.67**	**91.94**	**98.38**	**99.65**
LSUN-Resize	Ours	85.41	96.84	96.86	**88.49**	**97.56**	97.55
	MSP	47.45	91.30	98.11	16.54	75.32	94.03
	Mahalanobis	35.64	88.12	97.45	67.20	93.97	**98.70**
	Energy Score	71.75	94.12	98.64	21.38	79.29	94.97
	GEM	35.45	88.09	97.43	67.09	94.01	**98.70**
	VOS	**85.85**	**97.19**	**99.41**	63.91	90.95	97.74
iSUN	Ours	79.62	95.64	95.68	**82.15**	**96.05**	96.24
	MSP	43.40	89.72	97.72	17.02	75.87	94.20
	Mahalanobis	26.77	87.87	97.33	64.07	92.69	**98.32**
	Energy Score	66.27	92.56	98.25	19.20	78.98	94.90
	GEM	36.80	87.85	93.33	64.10	92.73	**98.32**
	VOS	**82.29**	**96.53**	**99.25**	63.01	90.91	97.95
Average Performance	Ours	**81.39**	**96.08**	96.17	**72.93**	**93.53**	93.69
	MSP	50.63	91.46	98.07	19.87	75.97	94.09
	Mahalanobis	46.83	90.46	97.92	48.35	84.90	96.37
	Energy Score	68.31	92.32	97.96	27.91	80.44	95.15
	GEM	50.81	90.45	97.91	48.36	84.94	96.38
	VOS	77.91	94.84	**98.60**	65.92	91.08	**98.03**

Table 6: Comparison in computation time and memory cost during inference.

Method	CIFAR-10 Time/Sample	CIFAR-10 Memory	CIFAR-100 Time/Sample	CIFAR-100 Memory
Ours	3.0ms	**1048.22MiB**	3.0ms	**1134.32MiB**
MSP	**0.02ms**	1825.21MiB	**0.02ms**	1825.98MiB
Mahala.	30.61ms	1983.17MiB	56.24ms	1983.81MiB
Energy	0.22ms	1830.01MiB	0.21ms	1838.23MiB
GEM	25.62ms	1983.51MiB	56.27ms	1984.77MiB
VOS	19.56ms	1984.05MiB	20.32ms	1984.91MiB

our approach. According to Fig. 4(b), $\max_g \frac{r_B - r_{A(g)}}{2r_B} |\mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g]|$ helps distinguish between ID samples and OOD samples after using the pretrained model and the contaminated dataset. Although the compared methods may show similar performance, their training time (e.g., PANDA) is much higher than ours.

Large-Scale Datasets: We evaluated our approach on large-scale datasets following a similar procedure. Specifically, we used the pretrained model CLIP (Radford et al., 2021) as the feature extractor, with ImageNet-1K (Huang et al., 2021) as the ID dataset and iNaturalist (Van Horn et al., 2018), SUN (Xiao et al., 2010), Textures (Cimpoi et al., 2014), and Places (Zhou et al., 2017) as the OOD datasets. The extracted text embeddings of ImageNet-1K labels served as the ID features for our OI-based detector. The utilized prompt

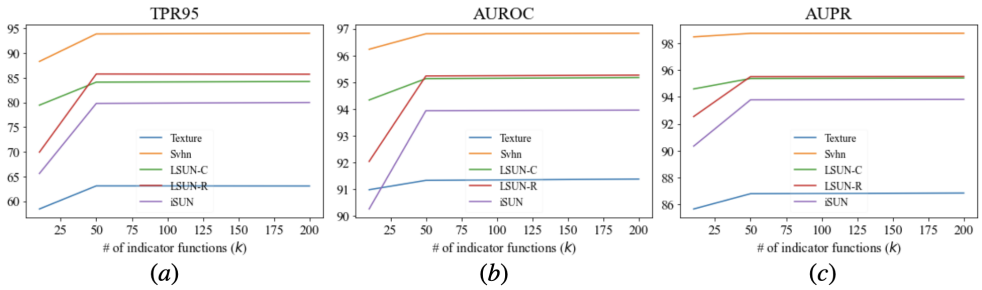

Figure 5: Performance of our approach with different numbers ($k = 10, 50, 200$) of condition functions for CIFAR-10 being ID data.

Table 7: AUROC on large-scale datasets. The ID dataset is ImageNet-1K. The backbone is ViT-B/16.

OOD Datasts	iNaturalist	SUN	Places	Textures	Ave.
MSP (Hendrycks & Gimpel, 2017)	87.44	79.73	79.67	79.69	81.633
ODIN (Liang et al., 2018)	94.65	87.17	85.54	87.85	88.803
GradNorm (Huang et al., 2021)	72.56	72.86	73.70	70.26	72.345
ViM (Wang et al., 2022b)	93.16	87.19	83.75	87.18	87.820
KNN	94.52	92.67	91.02	85.67	90.970
VOS (Du et al., 2022)	94.62	92.57	91.23	86.33	91.188
NPOS (Tao et al., 2022)	96.19	90.44	**89.44**	88.80	91.218
Ours	**99.57**	**92.78**	88.31	**94.92**	**93.895**

is "the < >" with the blank filled by specific labels. For example, if the class is cat, then the generated prompt would be "the cat". During inference, for a given image, we extracted its image embedding using CLIP and calculated the OI value between this embedding and the text embeddings of ID labels. To build the contaminated dataset, we merged all ID and OOD features, then randomly selected 100 features. During testing, each feature was subtracted from the mean vector of this contaminated dataset. The compared methods are allowed to use ImageNet-1K training data to train or fine-tune their OOD detectors. The results, shown in Table 7, demonstrate consistent performance of our approach on large-scale datasets.

Gaussian-Based Approaches: We feed raw data into WideResNet pretrained models to extract complex high-dimensional features whose shape is $128 \times 8 \times 8$. The pretrained models are downloaded from (Morteza & Li, 2022). The compared methods are MSP (Hendrycks & Gimpel, 2017), Mahalanobis (Lee et al., 2018), Energy score (Liu et al., 2020), GEM (Morteza & Li, 2022), and VOS (Du et al., 2022). All baseline methods are evaluated in the same feature space with their optimal hyperparameters for fair comparisons. We formed a small dataset by randomly selecting ten samples from each class to form the available ID dataset. Table 5 shows the experimental results of our approach compared with Gaussian-based approaches for the feature-level OOD detection when CIFAR-10 and CIFAR-100 are the ID datasets and Textures, SVHN, LSUN-Crop, LSUN-Resize, and iSUN are the OOD datasets. Table 6 shows the computation time and memory cost of our approach and the other methods. Our approach outperforms the other methods using the least memory, showing the highest TPR95 and AUROC on average. The AUPR of our approach is in the same range as other baseline methods. We further evaluated our approach using different numbers k of conditions functions and plotted the results in Fig. 5. From the figure, the performance of our approach increases with k and eventually converges, which is consistent with Fig. 3(b).

Backdoor Detection: We applied our approach to backdoor detection (Gu et al., 2019) by considering clean samples as ID data and poisoned samples as OOD data. We created different backdoored models and used them to extract data features. We compare with previous baseline methods plus STRIP (Gao et al., 2019). The average performance is given in Table 8. Our approach detects various attacks (Chen et al., 2017; Liu et al., 2019; Li et al., 2021b; Fu et al., 2022; Nguyen & Tran, 2021) and outperforms other baseline methods by showing the highest average AUROC. Details are provided in Appendix .8.

Table 8: Average performance of different OOD detectors for backdoor detection.

Metrics (%)	Ours	STRIP	Mahalanobis	GEM	MSP
TPR95	89.40	39.60	56.97	**91.57**	39.24
AUROC	**96.68**	70.30	75.94	58.08	54.92
AUPR	**95.42**	68.76	76.37	75.88	60.52

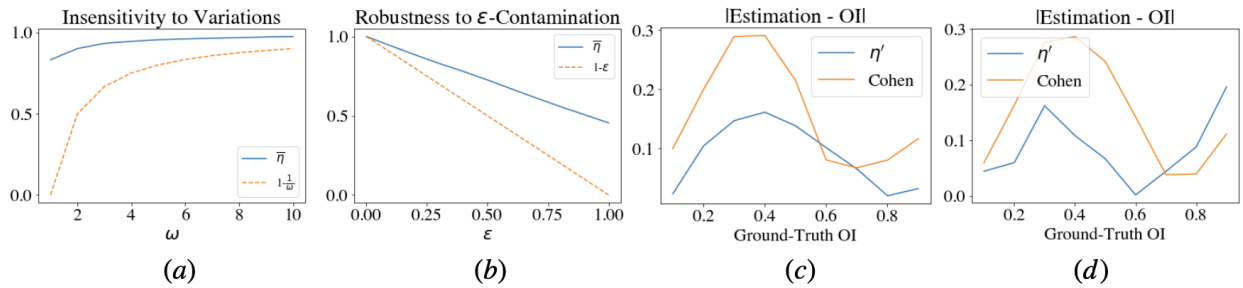

Figure 6: (a): illustration of **Proposition 5.1**. (b): illustration of **Proposition 5.2**. (c,d): the absolute value of estimation errors between the ground-truth OI and $\overline{\eta}'$ for uniform distributions (c) and truncated Gaussian distributions (d).

5 Analysis

5.1 Mathematical Properties

Proposition 5.1. *Consider uniform distribution for D^+ in $[0,1]$ and $1+\sin 2\pi\omega x$ for D^- in $[0,1]$, with $g(x) = \mathbb{1}\{||x|| \leq r\}$, the proposed confidence score function $\overline{\eta}$ is insensitive to small distribution variations:*

$$\overline{\eta}(D^+, D^-) \geq 1 - \frac{1}{\omega} \quad \forall\ \omega \geq 1. \tag{6}$$

Proposition 5.2. *Let $\epsilon \in [0,1]$, then for arbitrary D^+, D^-, and g, the proposed confidence score function $\overline{\eta}$ is robust against Huber-ϵ contamination:*

$$\overline{\eta}(D^+, (1-\epsilon)D^+ + \epsilon D^-) \geq 1 - \epsilon. \tag{7}$$

The insensitivity and robustness are illustrated in Fig. 6(a,b). The following theorem shows the potential of our OI-based confidence score function for estimating the model accuracy.

Theorem 5.3. *Assume that D and D^* are two different data distributions with $\eta(D, D^*) < 1$ and denote the overall accuracy of the model on D^* as Acc. If a model has a training accuracy p on D and a testing accuracy q on $D^* \setminus D$, then we have*

$$Acc \leq (p-q)(1 - \frac{1}{2r_B}||\mu_D - \mu_{D^*}|| - \max_g \frac{r_B - r_{A(g)}}{2r_B}\,|\mathbb{E}_D[g] - \mathbb{E}_{D^*}[g]|) + q. \tag{8}$$

Theorem 5.3 provides a theoretical interest for OOD generalization and is also empirically useful when the domain shift happens in the backdoor setting, where the model has zero clean accuracy on poisoned data (i.e., $q = 0$). Define the clean distribution as D, poisoned distribution as D^p, and a testing distribution D^* with $D^* = \sigma D + (1-\sigma)D^p$, where $\sigma \in [0,1]$. Then (8) becomes

$$Acc \leq p(1 - \frac{1-\sigma}{2r_B}||\mu_D - \mu_{D^p}|| - (1-\sigma)\max_g \frac{r_B - r_{A(g)}}{2r_B}|\mathbb{E}_D[g] - \mathbb{E}_{D^p}[g]|). \tag{9}$$

The experimental illustration is given in Appendix .9.

5.2 OI Estimation

Let $\overline{\eta}' = \max\{0, 1 - \max_g \frac{r_B - r_{A(g)}}{2r'} |\mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g]| - \frac{1}{2r'}||\mu_{D^+} - \mu_{D^-}||\}$ be the variant of $\overline{\eta}$ given in equation 6, where $r' \leq r_B$. For any given r', we could use Alg. 1 to calculate $\overline{\eta}'$ as an estimate of OI. To validate $\overline{\eta}'$, we consider estimate OI in $\mathbb{R}^4$. We set r' as the median Euclidean norm among all vectors in B, use indicator functions $g_j(x) = \mathbb{1}\{||x|| \leq r_j\}$, and choose $k = 100$ and $m = 50$. The utilization distributions are truncated Gaussian and uniform. We merged the two data clusterings and computed the mean vector of the combined set. We set this mean vector as the origin. The results are illustrated in Fig. 6(c,d). We empirically observed that the computation and memory costs of using the kernel estimator (Pastore & Calcagnì, 2019) in $\mathbb{R}^4$ are much higher than ours and Cohen's d measure (Inman & Bradley Jr, 1989). Therefore, we only show the comparison results with Cohen's d measure. The findings suggest that our OI-based confidence score function shows promise for applications in OI estimation.

6 Limitation and Broader Impact

Zhang et al. (2021a) proves that for any OOD detector, there exist situations where the OOD samples lie in the high probability or density regions of ID samples, and the OOD method performs no better than the random guess. We empirically observed that by shifting in- and out-distribution clusterings to the origin, our approach becomes less effective, supporting Zhang et al. (2021a)'s claim. Due to this limitation, the attacker may use our work to design stealthy OOD samples to evade our detection, causing machine learning systems to malfunction. Nevertheless, we believe that the positive impact of our approach far outweighs any potential negative societal impact.

7 Conclusion

This paper proposes a novel OOD detection approach that strikes a favorable balance between accuracy and computational efficiency. The utilized OI-based confidence score function is non-parametric, computationally lightweight, and demonstrates effective performance. Empirical evaluations indicate that the proposed OI-based OOD detector is competitive with state-of-the-art OOD detectors across a variety of scenarios, while maintaining a more frugal computational and memory footprint. The proposed OI-based confidence score function shows an insensitivity to small distributional shifts and a robustness against Huber ϵ-contamination. Additionally, it shows potential for estimating OI and model accuracy in specific applications. Overall, this paper showcases the effectiveness of using the OI-based metric for OOD detection.

References

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4183–4192, 2020.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 93–104, 2000.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Michael B Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. *Journal of the ACM*, 68(1):1–39, 2021.

Francesco De Comité, François Denis, Rémi Gilleron, and Fabien Letouzey. Positive and unlabeled examples help learning. In *Proceedings of International Conference on Algorithmic Learning Theory*, pp. 219–230, 1999.

Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *Proceedings of International Conference on Machine Learning*, pp. 685–693, 2014.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

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

Dheeru Dua and Casey Graff. UCI machine learning repository. "`http://archive.ics.uci.edu/ml`", 2017. Accessed: 2022-04-09.

Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(6), 2006.

Hao Fu, Akshaj Kumar Veldanda, Prashanth Krishnamurthy, Siddharth Garg, and Farshad Khorrami. A feature-based on-line detector to remove adversarial-backdoors by iterative demarcation. *IEEE Access*, 10: 5545 – 5558, 2022.

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the Annual Computer Security Applications Conference*, pp. 113–125, 2019.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Proceedings of Advances in Neural Information Processing Systems*, volume 31, pp. 1–9, 2018.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7:47230–47244, 2019.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *Proceedings of International Conference on Learning Representations*, 2017.

Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In *Proceedings of Advances in Neural Information Processing Systems*, pp. 677–689, 2021.

Peter J Huber. Robust estimation of a location parameter. *Breakthroughs in Statistics: Methodology and Distribution*, pp. 492–518, 1992.

Henry F Inman and Edwin L Bradley Jr. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of the overlap of two normal densities. *Communications in Statistics-theory and Methods*, 18(10):3851–3874, 1989.

Michael Kapralov, Vamsi Potluru, and David Woodruff. How to fake multiply by a gaussian matrix. In *Proceedings of International Conference on Machine Learning*, pp. 2101–2110, 2016.

Shehroz S Khan and Michael G Madden. One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(3):345–374, 2014.

Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4): 43–59, 2017.

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

Gukyeong Kwon, Mohit Prabhushankar, Dogancan Temel, and Ghassan AlRegib. Backpropagated gradient representations for anomaly detection. In *Proceedings of European Conference on Computer Vision*, pp. 206–226, 2020.

Nathaniel Lahn, Sharath Raghvendra, and Kaiyi Zhang. A combinatorial algorithm for approximating the optimal transport in the parallel and mpc settings. *Advances in Neural Information Processing Systems*, 36:21675–21686, 2023.

Larry Wasserman. Optimal transport and wasserstein distance. `https://www.stat.cmu.edu/%7elarry/=sml/Opt.pdf`, 2019a. Accessed: 2023-04-09.

Larry Wasserman. Lecture notes 27 36-705. `https://www.stat.cmu.edu/~larry/=stat705/Lecture27.pdf`, 2020b. Accessed: 2023-01-09.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Proceedings of Advances in Neural Information Processing Systems*, volume 31, pp. 1–9, 2018.

Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9664–9674, 2021a.

Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11578–11589, 2023.

Yuezun Li, Yiming Li, Baoyuan Wu, Longkang Li, Ran He, and Siwei Lyu. Invisible backdoor attack with sample-specific triggers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16463–16472, 2021b.

Zheng Li, Yue Zhao, Xiyang Hu, Nicola Botta, Cezar Ionescu, and George Chen. Ecod: Unsupervised outlier detection using empirical cumulative distribution functions. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

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

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *Proceedings of IEEE International Conference on Data Mining*, pp. 413–422, 2008.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In *Proceedings of Advances in Neural Information Processing Systems*, volume 33, pp. 21464–21475, 2020.

Yingqi Liu, Wen-Chuan Lee, Guanhong Tao, Shiqing Ma, Yousra Aafer, and Xiangyu Zhang. Abs: Scanning neural networks for back-doors by artificial brain stimulation. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1265–1282, 2019.

David Macêdo, Tsang Ing Ren, Cleber Zanchettin, Adriano LI Oliveira, and Teresa Ludermir. Entropic out-of-distribution detection: Seamless detection of unknown examples. *IEEE Transactions on Neural Networks and Learning Systems*, 33(6):2350–2364, 2021.

Alireza Makhzani and Brendan J Frey. Winner-take-all autoencoders. In *Proceedings of Advances in Neural Information Processing Systems*, volume 28, pp. 2791–2799, 2015.

Andrey Malinin and Mark Gales. Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *Proceedings of Advances in Neural Information Processing Systems*, volume 32, pp. 1–9, 2019.

Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 8, 2022.

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

Tuan Anh Nguyen and Anh Tuan Tran. Wanet - imperceptible warping-based backdoor attack. In *Proceedings of the International Conference on Learning Representations*, 2021.

Massimiliano Pastore and Antonio Calcagnì. Measuring distribution similarities between samples: A distribution-free overlapping index. *Frontiers in Psychology*, 10:1089, 2019.

Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2898–2906, 2019.

Abhijeet Phatak, Sharath Raghvendra, Chittaranjan Tripathy, and Kaiyi Zhang. Computing all optimal partial transports. In *International Conference on Learning Representations*, 2023.

Chen Qiu, Aodong Li, Marius Kloft, Maja Rudolph, and Stephan Mandt. Latent outlier exposure for anomaly detection with contaminated data. In *Proceedings of International Conference on Machine Learning*, pp. 18153–18167, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning*, pp. 8748–8763, 2021.

Sharath Raghvendra, Pouyan Shirzadian, and Kaiyi Zhang. A new robust partial p-wasserstein-based metric for comparing distributions. *arXiv preprint arXiv:2405.03664*, 2024.

Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2806–2814, 2021.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14318–14328, 2022.

Peter J Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *Proceedings of International Conference on Machine Learning*, pp. 4393–4402, 2018.

Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *Proceedings of International Conference on Learning Representations*, 2020.

Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14902–14912, 2021.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *Proceedings of International Conference on Information Processing in Medical Imaging*, pp. 146–157, 2017.

Friedrich Schmid and Axel Schmidt. Nonparametric estimation of the coefficient of overlapping—theory and empirical application. *Computational Statistics & Data Analysis*, 50(6):1583–1596, 2006.

Sarah Schneider, Doris Antensteiner, Daniel Soukup, and Matthias Scheutz. Autoencoders-a comparative analysis in the realm of anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1986–1992, 2022.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *Proceedings of International Joint Conference on Neural Networks*, pp. 1453–1460, 2011.

Leitian Tao, Xuefeng Du, Jerry Zhu, and Yixuan Li. Non-parametric outlier synthesis. In *Proceedings of The International Conference on Learning Representations*, 2022.

David Martinus Johannes Tax. *One-class classification: Concept learning in the absence of counter-examples*. PhD thesis, Technische Universiteit Delft, 2002.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Chenyang Wang, Junjun Jiang, Xiong Zhou, and Xianming Liu. Resmooth: Detecting and utilizing ood samples when training with data augmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022a.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022b.

Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 529–534, 2011.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010.

Hongzuo Xu, Guansong Pang, Yijie Wang, and Yongjun Wang. Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Pingmei Xu, Krista A Ehinger, Yinda Zhang, Adam Finkelstein, Sanjeev R Kulkarni, and Jianxiong Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Yannis G Yatracos. Limitations of the wasserstein mde for univariate data. *Statistics and Computing*, 32(6): 95, 2022.

Zhiyuan You, Kai Yang, Wenhan Luo, Lei Cui, Yu Zheng, and Xinyi Le. Adtr: Anomaly detection transformer with feature reconstruction. In *Proceedings of Neural Information Processing*, pp. 298–310, 2023.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Adriano Z Zambom and Dias Ronaldo. A review of kernel density estimation with applications to econometrics. *International Econometric Review*, 5(1):20–42, 2013.

Lily Zhang, Mark Goldstein, and Rajesh Ranganath. Understanding failures in out-of-distribution detection with deep generative models. In *Proceedings of International Conference on Machine Learning*, pp. 12427–12436, 2021a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Yivan Zhang, Gang Niu, and Masashi Sugiyama. Learning noise transition matrix from only noisy labels via total variation regularization. In *Proceedings of International Conference on Machine Learning*, pp. 12501–12512, 2021b.

Yufeng Zhang, Jialu Pan, Wanwei Liu, Zhenbang Chen, Kenli Li, Ji Wang, Zhiming Liu, and Hongmei Wei. Kullback-leibler divergence-based out-of-distribution detection with flow-based generative models. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Zhilin Zhao, Longbing Cao, and Kun-Yu Lin. Out-of-distribution detection by cross-class vicinity distribution of in-distribution data. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464, 2017.

Alon Zolfi, Guy Amit, Amit Baras, Satoru Koda, Ikuya Morikawa, Yuval Elovici, and Asaf Shabtai. Yolood: Utilizing object detection concepts for multi-label out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2024.

.1 Proof of Theorem 3.3

Definition .1 (Total Variation Distance (TVD))**.** TVD $\delta : \mathbb{R}^n \times \mathbb{R}^n \to [0, 1]$ of P and Q is defined as:

$$\delta(P, Q) = \frac{1}{2} \int_{\mathbb{R}^n} |f_P(x) - f_Q(x)| \, dx. \tag{10}$$

Since OI + TVD = 1, we have $\eta = 1 - \delta = 1 - \delta_A - \delta_{\mathbb{R}^n \setminus A}$.

Proof. Let f_{D^+} and f_{D^-} be the probability density functions for D^+ and D^-. From **Definition** .1, we have

$$\eta(D^+, D^-) = 1 - \delta_A(D^+, D^-) - \delta_{A^c}(D^+, D^-). \tag{11}$$

Using (11), triangular inequality, and boundedness, we obtain

$$\begin{aligned}
||\mu_{D^+} - \mu_{D^-}|| &= || \int_B x \left(f_{D^+}(x) - f_{D^-}(x) \right) dx|| && (12)\\
&\leq \int_B ||x(f_{D^+}(x) - f_{D^-}(x))||dx && (13)\\
&= \int_A ||x|| \cdot |f_{D^+}(x) - f_{D^-}(x)|dx && (14)\\
&\quad + \int_{A^c} ||x|| \cdot |f_{D^+}(x) - f_{D^-}(x)|dx && (15)\\
&\leq 2r_A\delta_A + 2r_{A^c}\delta_{A^c} && (16)\\
&= 2r_A\delta_A + 2r_{A^c}(1 - \delta_A - \eta(D^+, D^-)) && (17)
\end{aligned}$$

which implies (3). Since $1 - \delta_A - \eta(D^+, D^-) \geq 0$, we can replace r_{A^c} with r_B in (17) to get (4). □

.2 Proof of Corollary 3.4

Proof. Let $g : B \to \{0, 1\}$ be a condition function and define $A(g) = \{x \mid g(x) = 1, x \in B\}$. According to the definition of δ_A and triangular inequality, we have

$$\delta_{A(g)}(D^+, D^-) = \frac{1}{2} \int_{A(g)} |f_{D^+}(x) - f_{D^-}(x)| dx \tag{18}$$

$$\geq \frac{1}{2} | \int_{A(g)} f_{D^+}(x) - f_{D^-}(x) dx | \tag{19}$$

$$= \frac{1}{2} | \int_{\mathbb{R}^n} g(x) f_{D^+}(x) - g(x) f_{D^-}(x) dx |$$

$$= \frac{1}{2} \left| \mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g] \right| . \tag{20}$$

Applying (20) into **Theorem** 3.3 gives **Corollary** 3.4. □

.3 Proof of Proposition 5.1

Proof. Since the support is $[0, 1]$, we have $r_B = 1$, $\mu_{D^+} - \mu_{D^-} = -\int_0^1 x \sin 2\pi\omega x dx \leq \frac{1}{\omega}$, $\mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g] = -\int_0^1 g(x) \sin 2\pi\omega x dx \leq \frac{1}{\omega}$, and $\overline{\eta} \geq 1 - \frac{1}{2\omega} - \max_g \frac{1 - r_{A(g)}}{2} \frac{1}{\omega} \geq 1 - \frac{1}{\omega}$. □

.4 Proof of Proposition 5.2

Proof. Denote $Q = (1 - \epsilon) D^+ + \epsilon D^-$, then $\mu_{D^+} - \mu_Q = \epsilon(\mu_{D^+} - \mu_{D^-})$, $\mathbb{E}_{D^+}[g] - \mathbb{E}_Q[g] = \epsilon(\mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g])$, and $\overline{\eta}(D^+, Q) = 1 - \epsilon(\frac{1}{2r_B} ||\mu_{D^+} - \mu_{D^-}|| + \max_g \frac{r_B - r_{A(g)}}{2r_B} |\mathbb{E}_{D^+}[g] - \mathbb{E}_{D^-}[g]|) \geq 1 - \epsilon$. □

.5 Proof of Theorem 5.3

Proof. Let f_D and f_{D^*} be their probability density functions, $Acc = \int_{x \sim D^*} \left(p \frac{\min\{f_D(x), f_{D^*}(x)\}}{f_{D^*}(x)} + q \left(1 - \frac{\min\{f_D(x), f_{D^*}(x)\}}{f_{D^*}(x)} \right) \right) \times f_{D^*}(x) dx = p\eta(D, D^*) + q(1 - \eta(D, D^*)) \leq (p - q)(1 - \frac{1}{2r_B} ||\mu_D - \mu_{D^*}|| - \max_g \frac{r_B - r_{A(g)}}{2r_B} |\mathbb{E}_D[g] - \mathbb{E}_{D^*}[g]|) + q$. □

.6 Details of Used Datasets

Details of used datasets are given in Table 9.

Table 9: Information on other utilized datasets.

Dataset	Dimension	# Class	# Training	# Testing
CIFAR-10	$3 \times 32 \times 32$	10	50000	10000
CIFAR-100	$3 \times 32 \times 32$	100	50000	10000
Textures	$3 \times 32 \times 32$	47	N/A	5640
SVHN	$3 \times 32 \times 32$	10	N/A	26032
LSUN_C	$3 \times 36 \times 36$	1	N/A	10000
LSUN_R	$3 \times 32 \times 32$	1	N/A	10000
iSUN	$3 \times 32 \times 32$	1	N/A	8925
MNIST	$1 \times 28 \times 28$	10	60000	10000
GTSRB	$3 \times 32 \times 32$	43	39209	12630
YouTube Face	$3 \times 55 \times 47$	1283	103923	12830
sub-ImageNet	$3 \times 224 \times 224$	200	100000	2000

.7 Details for Fig. 2

The numerical results are given in Table 10.

Table 10: AUROC (%) for different methods on UCI datasets.

Ours 97.67 ± 2.55	L1-Ball 90.19 ± 22.21	K-Center 92.4 ± 10.24	Parzen 93.99 ± 9.92
Gaussian 95.83 ± 6.45	K-Mean 94.51 ± 5.7	1-Nearest Neighbor 93.24 ± 11.05	K-Nearest Neighbor 93.24 ± 11.05
Auto-Encoder Network 87.81 ± 22.15	Linear Programming 81.5 ± 34.42	Principal Component 81.8 ± 25.18	Lof Range 86.04 ± 23.54
Nearest Neighbor Distance 84.54 ± 15.91	Minimum Spanning Tree 92.32 ± 11.4	Minimum Covariance Determinant 94.1 ± 11.56	Self Organizing Map 93.6 ± 7.49
Support Vector Machine 78.69 ± 36.72	Minimax Probability Machine 77.84 ± 35.57	Mixture Gaussians 93.61 ± 11.15	Local Outlier Factor 89.04 ± 13.77
Naive Parzen 97.41 ± 2.59	Local Correlation Integral 96.79 ± 2.45		

.8 Backdoor Detection

The datasets are balanced by having an equal number of clean and poisoned samples. For each backdoor attack, we assume that a small clean validation dataset is available (i.e., ten samples from each class) at the beginning. Therefore, the poisoned samples (i.e., samples attached with triggers) can be considered OOD, whereas the clean samples can be considered ID. The metrics used are: TPR95 (i.e., the detection accuracy for poisoned samples when the detection accuracy for clean samples is 95%), AUROC, and AUPR. We have carefully fine-tuned the baseline methods' hyperparameters to ensure their best performance over other hyperparameter choices. Fig. 7 shows the utilized triggers, and Table 11 shows the details for backdoor detection. For most triggers, our method has over 96% of TPR95, over 97% of AUROC, and 95% of AUPR. Our detector is robust against the latest or advanced backdoor attacks, such as Wanet, invisible trigger, all label attack, and filter attack, whereas the baseline methods show low performance on those attacks. The Gaussian-based baseline methods encountered an error for two cases because the data features are dependent.

.9 Experimental Illustration of Theorem 5.3

Setup: We use MNIST, GTSRB, YouTube Face, and sub-ImageNet and their domain-shifted versions given in Fig. 8 to illustrate the theorem. We set $\sigma = 0, 0.1, ..., 0.9, 1$ and calculated the RHS of (9) using L_1, L_2, and L_∞ norms in the raw image input space, model output space, and hidden layer space. We use $g_j(x) = \mathbb{1}\{||x|| \leq r_j\}$. Fig. 9 shows the model accuracy and corresponding upper bounds with different norms and σs. The model accuracy is below all calculated upper limits, validating the theorem. The difference between model accuracy and the calculated upper bound accuracy can reflect the extent of the domain shift in the test dataset. From Fig. 9, a large difference reflects a large domain shift. When the domain shift vanishes, the model accuracy and the calculated upper bound accuracy are close.

From Fig. 9, the calculated upper bound accuracy varies with spaces. The inference is that a low calculated upper bound accuracy implies a high likelihood of detecting domain-shifted data in that particular utilized space. For example, in Fig. 9 MNIST and YouTube, the input space with L_∞ norm shows the lowest upper

Table 11: Comparison results for backdoor detection (higher number implies higher accuracy).

Datasets	Trigger	Metrics (%)	Ours	STRIP	Mahalanobis	GEM	MSP
MNIST	All label	TPR95	83.05	2.58	50.83	**100**	100
		AUROC	**96.13**	44.69	90.78	50.43	50
		AUPR	**94.20**	35.47	86.71	70.94	70.83
MNIST	Naive.1	TPR95	**100**	98.85	99.86	100	5.11
		AUROC	**97.50**	97.32	97.49	53.95	51.64
		AUPR	96.17	95.95	**96.38**	74.74	50.41
MNIST	Naive.2	TPR95	96.53	67.46	35.16	**100**	14.69
		AUROC	**97.28**	93.67	78.63	53.51	58.14
		AUPR	**95.75**	89.85	78.65	74.62	64.16
CIFAR-10	TCA.1	TPR95	**100**	35.68	100	100	4.38
		AUROC	**97.50**	83.00	97.49	50	49.23
		AUPR	95.47	73.22	**97.84**	76.32	52.64
CIFAR-10	TCA.2	TPR95	**100**	27.86	100	100	0.02
		AUROC	**97.50**	68.79	97.49	50	29.90
		AUPR	**97.63**	72.41	95.86	67.86	18.05
CIFAR-10	Wanet	TPR95	37.87	0.07	20.35	22.90	**100**
		AUROC	**92.74**	34.97	50.61	57.81	50
		AUPR	**89.95**	37.42	57.30	68.48	74.87
GTSRB	Moving	TPR95	**99.99**	54	Fail: dependent data features		
		AUROC	**85.39**	7.29			
		AUPR	**96.96**	89.07			
GTSRB	Filter	TPR95	**85.39**	7.29	Fail: dependent data features		
		AUROC	**96.54**	38.92			
		AUPR	**95.42**	38.81			
GTSRB	Wanet	TPR95	**100**	1.24	0.51	100	100
		AUROC	**97.50**	36.31	54.46	50	50
		AUPR	**97.62**	39.53	48.92	75.23	75.23
YouTube Face	Sunglasses	TPR95	73.37	83.03	71.64	**98.58**	13.06
		AUROC	**95.21**	94.80	94.38	84.29	66.55
		AUPR	93.00	**95.54**	94.63	88.83	53.27
YouTube Face	Lipstick	TPR95	**96.64**	90.14	90.88	94.18	3.73
		AUROC	**97.21**	93.15	93.26	80.80	50.14
		AUPR	**96.30**	94.98	95.09	86.53	53.27
sub-ImageNet	Invisible	TPR95	**100**	7.01	0.5	100	51.40
		AUROC	**97.49**	66.26	4.78	50	93.61
		AUPR	**96.53**	62.83	12.27	75.26	92.46
Average Performance		TPR95	89.40	39.60	56.97	**91.57**	39.24
		AUROC	**96.68**	70.30	75.94	58.08	54.92
		AUPR	**95.42**	68.76	76.37	75.88	60.52

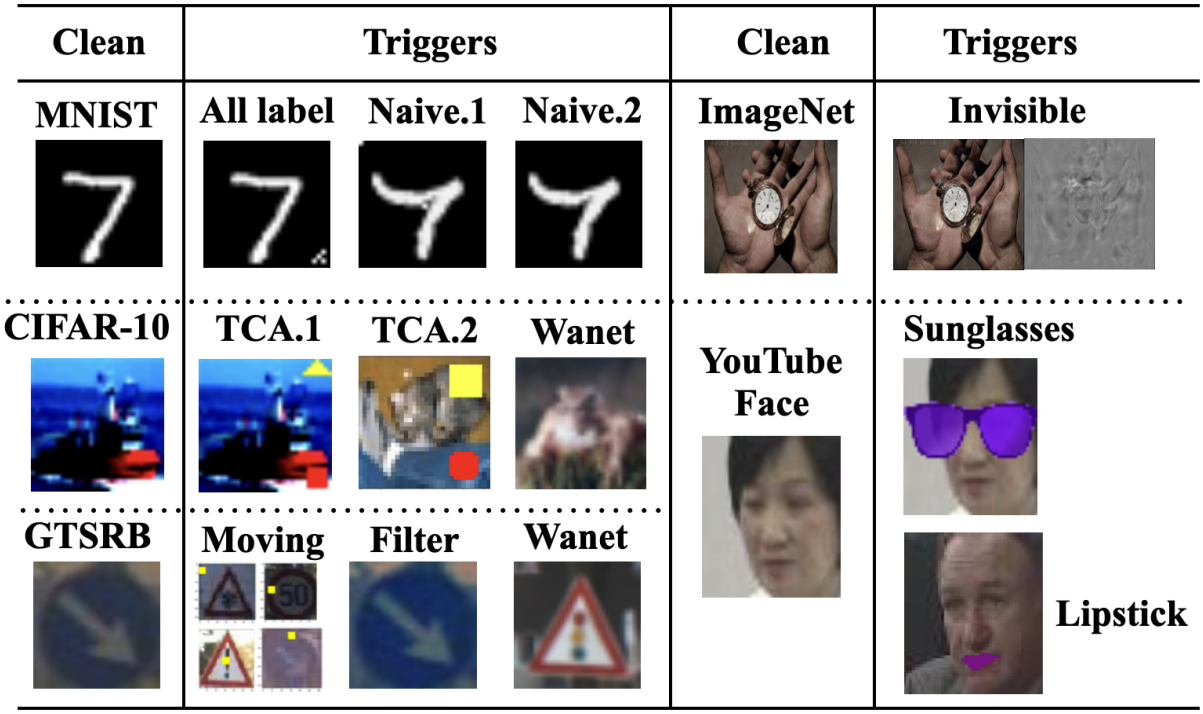

Figure 7: Pictures under "Triggers" are poisoned samples. Pictures under "Clean" are clean samples.

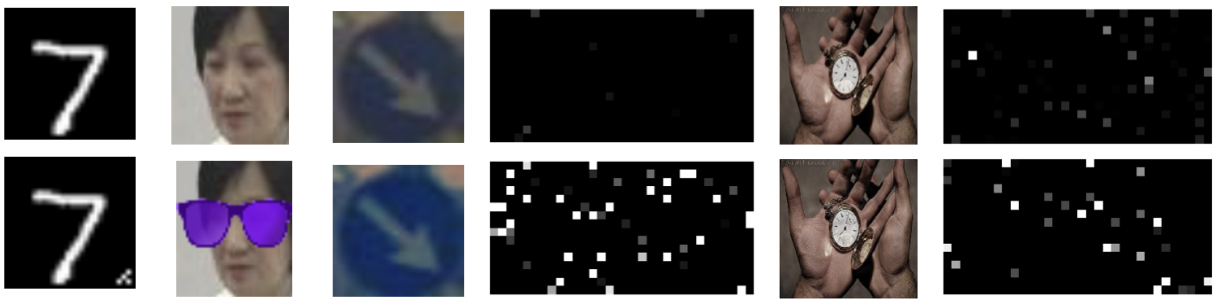

Figure 8: Top: original samples. Bottom: domain-shifted samples. From left to right: MNIST, YouTube Face, GTSRB, GTSRB in hidden layer space, ImageNet, ImageNet in hidden layer space.

bound accuracy. Therefore, domain-shifted samples will likely be detected in the input space. Indeed, from Fig. 8, even vision inspection can easily detect them. As for GTSRB and ImageNet, the input space has the highest upper limit lines. Therefore, domain-shifted samples are less likely to be detected in the input space. Fig. 8 shows that the visual inspection can barely detect them. However, the hidden layer space gives the lowest upper bound accuracy. Therefore, domain-shifted samples are more likely to be distinguished in the hidden layer space, as shown in Fig. 8. The AUROC is 96.54% for GTSRB and 97.49% for ImageNet according to Table 11.

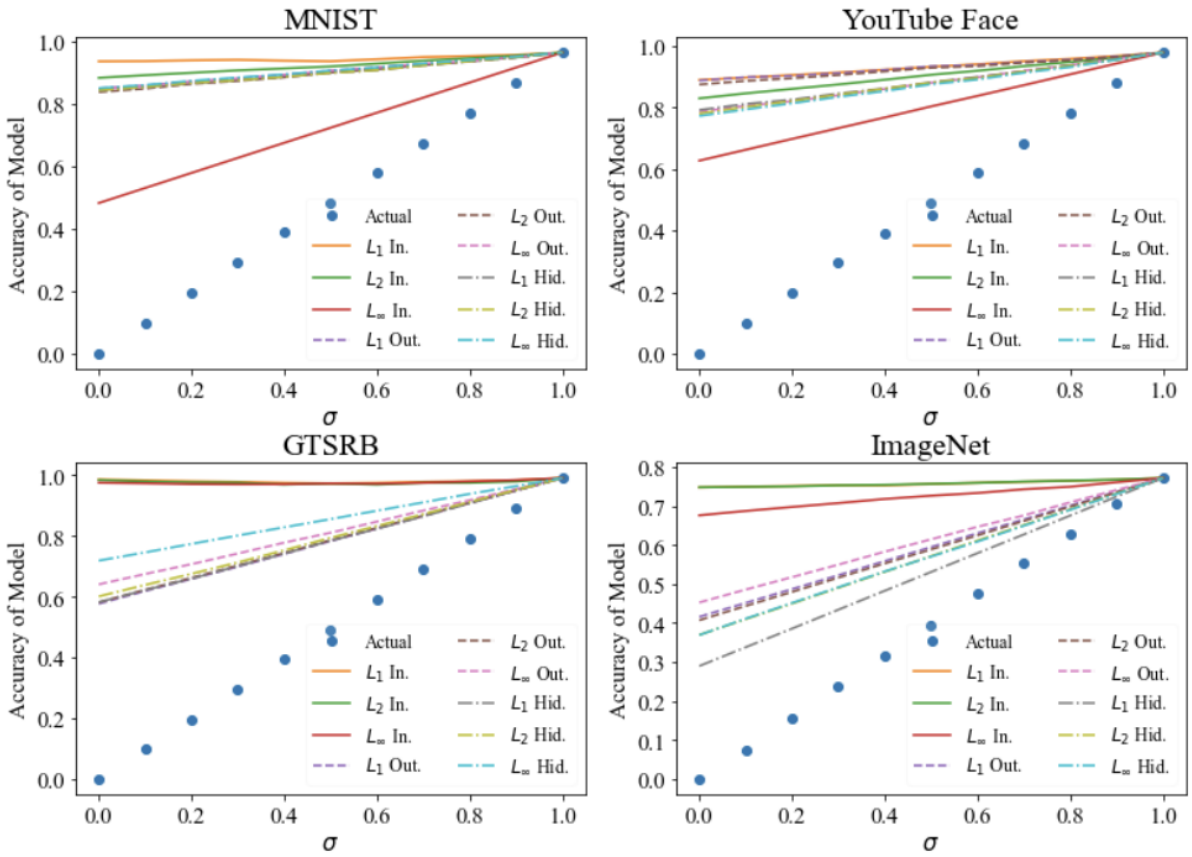

Figure 9: The model accuracy vs. equation 9 calculated with $L_{1,2,\infty}$ norms in input, output, and hidden spaces. X-axis is the ratio of clean samples to the entire testing samples.