BOOSTING OUT-OF-DISTRIBUTION DETECTION WITH MULTIPLE PRE-TRAINED MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Out-of-Distribution (OOD) detection, i.e., identifying whether an input is sampled from a novel distribution other than the training distribution, is a critical task for safely deploying machine learning systems in the open world. Recently, post hoc detection utilizing pre-trained models has shown promising performance and can be scaled to large-scale problems. This advance raises a natural question: Can we leverage the diversity of multiple pre-trained models to improve the performance of post hoc detection methods? In this work, we propose a detection enhancement method by ensembling multiple detection decisions derived from a zoo of pre-trained models. Our approach uses the p-value instead of the commonly used hard threshold and leverages a fundamental framework of multiple hypothesis testing to control the true positive rate of In-Distribution (ID) data. We focus on the usage of model zoos and provide systematic empirical comparisons with current state-of-the-art methods on various OOD detection benchmarks. The proposed ensemble scheme shows consistent improvement compared to single-model detectors and significantly outperforms the current competitive methods. Our method substantially improves the relative performance by 65.40% and 26.96% on the CIFAR10 and ImageNet benchmarks.

1 INTRODUCTION

Deep neural networks have achieved empirical success in many applications, but generalization robustness has always been a thorny problem in deep learning. A sophisticated and well-trained deep neural network can provide excellent test performance on identically distributed (ID) test data but may fail to make accurate predictions on inputs from outside the training distribution Nguyen et al. (2015). This poses a big obstacle to the generalization of deep neural network models. Especially in safety-critical applications, it is better to identify out-of-distribution (OOD) inputs ahead of time rather than letting the model make predictions that may be unreliable.

On the basis of pre-trained deep neural networks, many recent works on post hoc OOD detection have proposed diverse score functions to distinguish OOD samples utilizing the output probability, logits, gradients, and features of the pre-trained classifier. At the same time, some works also propose new training strategies to encourage the network to learn more features that may not be relevant to the OOD classification task. For example, MSP (Hendrycks & Gimpel, 2017) uses the maximum softmax probability, Energy score (Liu et al., 2020) considers the logits, and GradNorm (Huang et al., 2021) employs the vector norm of gradients. Based on these frameworks, several improved methods such as ODIN (Liang et al., 2018), Adjusted Energy Score (Lin et al., 2021), ReAct (Sun et al., 2021) are proposed to enhance the performance of OOD detection. These score functions above measure the similarity between a test input and the training (ID) data through pre-trained feature extractors or classifiers. There are also many distance-based algorithms that directly quantify the distance of samples in the embedding space extracted from a pre-trained model and regard a test input as an OOD sample when it is far from the ID data. Lee et al. (2018) assumes the conditional distribution of features given the class label is a Gaussian distribution and derives a confidence score based on the Mahalanobis distance. SSD (Sehwag et al., 2020) considers self-supervised pre-training and a Mahalanobis distance. Tack et al. (2020) uses contrastive learning with distributionally-shifted augmentations for pre-training and proposes a detection score specific to their training scheme. Sun et al. (2022) studies the nearest-neighbor distance and demonstrates the efficacy of non-parametric modeling of the feature distribution for OOD detection tasks.

The performance of post hoc detection highly depends on the quality of pre-training. The most commonly used model architectures in OOD detection include convolutional networks such as ResNet (He et al., 2016), DenseNet (Huang et al., 2017) and Wide-ResNet (Zagoruyko & Komodakis, 2016), and of course Transformer models such as Swin (Liu et al., 2022) or ViT (Dosovitskiy et al., 2021). In general, the pre-trained models focus on the features related to classification tasks and the learnt representation may be insufficiently rich for OOD detection. Therefore, researchers have proposed ideas such as contrastive learning (Winkens et al., 2020; Tack et al., 2020), adversarial training Biggio & Roli (2018); Miller et al. (2020); Chalapathy & Chawla (2019) , outlier exposure (Hendrycks et al., 2018; Papadopoulos et al., 2021) or other auxiliary artificially synthesized data (Lee et al., 2017) and auxiliary loss function (Vyas et al., 2018) to encourage models to learn high-level, task-agnostic and comprehensive features, which makes the model more robust and efficient in the downstream detection task. These models trained with different architectures and training strategies can extract diverse features that may complement each other well. So, a natural question is raised:

Can we leverage the diversity of multiple pre-trained models to improve the performance of post hoc OOD detectors?

To answer this question, we first build a model zoo that captures as many properties of the input as possible and remains sensitive to distributional changes. Then we reformulate the OOD detection task to check whether there exists a model in the model zoo that can identify the test input as an OOD sample. Section 3.1 shows that the naive ensemble of multiple OOD detection decisions cannot maintain the true positive rate of the ID data (TPR). Therefore, we propose an ensemble scheme to integrate the results of multiple OOD detectors and provide theoretical guarantees that our method can keep TPR at the target level. In Section 4, we also report the empirical TPR of our method, which is close to the target TPR level.

Ensembling is not new to OOD detection. Morningstar et al. (2021) combines multiple test statistics from generative models to differentiate ID and OOD data. Haroush et al. (2022) uses both the Simes' method and Fisher's method to summarize p-values computed for each channel and layer of a deep neural network. Bergamin et al. (2022) shows that combining different types of test statistics using Fisher's method overall leads to a more accurate out-of-distribution test. Recently, Magesh et al. (2022) proposes an ensemble framework that combines any number of different test statistics using the Benjamini–Yekutieli procedure (Benjamini & Yekutieli, 2001) and a conformal p-value estimator (Vovk et al., 1999). In this work, we develop a simple and fundamental ensemble scheme for using model zoos in OOD detection and name our method **Z**oo-based **O**OD **D**etection **E**nhancement (**ZODE**). Our method directly estimates the p-values according to its definition and employs the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995) to control TPR. Then, we provide theoretical guarantees and empirical validation to show that ZODE can maintain the TPR close to its target level. On the other hand, we focus on the settings of the model zoo and conduct systematic experiments to demonstrate the superiority of our approach. First, we show that ZODE can consistently improve current OOD detectors. Second, by comparing single-model detectors with the ZODE-ensembled detector, we find that ZODE can exploit the diversity of multiple pre-trained models and leverage complementarity among single-model detectors. Finally, our approach significantly improves current SOTA performance.

We summarize our contributions as follows:

- We provide novel insights into OOD detection from the perspective of the model zoo. We propose an enhancement scheme, ZODE, for OOD detection by exploiting the diversity of pre-trained models. The proposed method is inspired by a simple and fundamental framework of multiple hypothesis testing. Our theoretical results and experiments clearly show that ZODE can leverage the complementarity among single-model detectors to improve performance.
- We point out that the naive ensemble of multiple OOD detectors leads to lower TPR. Then we provide theoretical analysis and empirical validation to demonstrate that our proposed method can maintain TPR well under the settings of the model zoo.
- Extensive experiments show that our method can effectively and consistently improve the power of identifying OOD samples. On a commonly used CIFAR10 benchmark, our method significantly improves the SOTA result of the average false positive rate from 11.07% to 3.83%. For a challenging OOD detection task based on ImageNet, we show

that our method is scalable to large-scale problems and significantly improves the SOTA result of the average false positive rate from 38.47% to 28.10%.

2 PRELIMINARIES

Out-of-Distribution Detection aims to check whether a given input is generated from the training distribution or not. It is a one-sample hypothesis testing problem if we can only access the training data. We denote $\mathcal{X}$ and $\mathcal{Y}$ as the input and label space respectively and let $\mathcal{P}_{id}$ be the training distribution over $\mathcal{X} \times \mathcal{Y}$. Suppose that $\phi(\mathbf{x})$ is a neural network trained on data drawn from $\mathcal{P}_{id}$ to predict the label of input $\mathbf{x} \in \mathcal{X}$. Let $\mathcal{D}_{id}$ denote the marginal distribution on $\mathcal{X}$. Then we call $\mathbf{x} \sim \mathcal{D}_{id}$ an in-distribution (ID) sample, otherwise we identify it as an "unknown" input, called out-of-distribution (OOD) data.

At test time, OOD detection distinguishes OOD data and ID data by using a decision function:

$$G(\mathbf{x}^*) = \begin{cases} ID & S(\mathbf{x}^*) \geq \lambda; \\ OOD & S(\mathbf{x}^*) < \lambda; \end{cases} \tag{1}$$

where $\mathbf{x}^*$ is a test input, $S(\cdot)$ is a score function to be defined that gives higher scores for ID and lower for OOD, and λ is the threshold. In this work, we consider post hoc OOD detection in which the score function S is derived from a pre-trained classifier ϕ, i.e. $S(\mathbf{x}^*) = S(\mathbf{x}^*; \phi)$.

We denote $F(s; \phi)$ as the distribution of $S(\mathbf{x}; \phi)$ with $\mathbf{x} \sim \mathcal{D}_{id}$ and any pre-trained model ϕ. Then, if $\mathbf{x}^*$ is an ID sample, the score $S(\mathbf{x}^*; \phi)$ is an ID value following the distribution $F(s; \phi)$. Therefore, given a pre-trained model zoo $\mathcal{M} = \{\phi_1, \ldots, \phi_m\}$, we strengthen the OOD detection problem to:

Is there $\phi \in \mathcal{M}$ that would allow us to identify $\mathbf{x}^*$ as an OOD sample based on $S(\mathbf{x}^*; \phi)$?

In this work, we proposed an approach to achieve the goal of this OOD detection problem.

3 METHODOLOGY

3.1 NAIVE ENSEMBLE CANNOT MAINTAIN TPR

To leverage the model zoo $\mathcal{M}$ for OOD detection, a straightforward way is to execute the detection procedure in Eq.(1) based on each pre-trained model:

$$G(\mathbf{x}^*; \phi) = \begin{cases} ID & S(\mathbf{x}^*; \phi) \geq \lambda_\phi; \\ OOD & S(\mathbf{x}^*; \phi) < \lambda_\phi; \end{cases} \tag{2}$$

and identify $\mathbf{x}^*$ as an OOD sample if there exists $\phi \in \mathcal{M}$ such that $G(\mathbf{x}^*; \phi) = OOD$, i.e.,

$$G(\mathbf{x}^*; \mathcal{M}) = \begin{cases} ID & \text{if } S(\mathbf{x}^*; \phi) \geq \lambda_\phi, \ \forall \phi \in \mathcal{M}; \\ OOD & \text{if } S(\mathbf{x}^*; \phi) < \lambda_\phi, \ \exists \phi \in \mathcal{M}; \end{cases} \tag{3}$$

In other words, $\mathbf{x}^*$ is classified as an ID sample only if all detectors $G(\mathbf{x}^*; \phi_i)$, $\phi_i \in \mathcal{M}$ agree that $\mathbf{x}^*$ is an ID sample. However, this simple approach is not easy to control the true positive rate of the ID data (TPR). In practice, the threshold λ_ϕ is chosen so that a high fraction (e.g. 95%) of ID data is correctly identified. We denote the target level of the true positive rate of the ID data as TPR_0 and write $\alpha = 1 - \text{TPR}_0$. Therefore, each detector $G(\mathbf{x}^*; \phi_i)$ has a α probability of misidentifying an ID sample as an OOD sample. When ensembling multiple single-model detectors, the probability of making mistakes also accumulates. It is easy to see that **the detector $G(\mathbf{x}^*; \mathcal{M})$ can misidentify an ID sample as an OOD sample with probability more than α, specifically $1 - (1-\alpha)^m$ when detectors are independent.** As more and more pre-trained models become available, this error probability of $G(\mathbf{x}^*; \mathcal{M})$ increases until it becomes 100%. This implies that the naive ensembled detector cannot maintain the target TPR level. On the other hand, by fixing m, we can assign a high probability to TPR_0 to make sure $1 - \text{TPR}_0^m = 5\%$. In this case, TPR_0 should be very large and even close to 1. This greatly reduces the probability of successfully identifying OOD data, as each single-model detector becomes very conservative and can only identify extreme OOD data. In this work, we develop an ensemble scheme that can maintain the target TPR level while keeping a high probability of successfully identifying OOD data.

3.2 USING P-VALUE FOR OOD DETECTION

However, directly integrating the score functions is uninterpretable and lacks theoretical guarantees. Therefore, we use the p-value for OOD detection. P-value (Abramovich & Ritov, 2013) is defined in the framework of statistical hypothesis testing. In OOD detection, the p-value is a probability measure that quantifies how extreme the observed score is when the input is an ID sample. For example, we identify an input $\mathbf{x}$ as an OOD sample (reject the null hypothesis) when the observed detection score $S(\mathbf{x})$ is smaller than a critical value γ. Given a test sample $\mathbf{x}^*$, the lower value of $S(\mathbf{x}^*)$, the more likely $\mathbf{x}^*$ is not drawn from the training distribution. Hence, the p-value of $\mathbf{x}^*$ is the probability that $S(\mathbf{x})$ is less than $S(\mathbf{x}^*)$ under the ID distribution, that is,

$$\text{P-value of } \mathbf{x}^* = P\big(S(\mathbf{x}) \leq S(\mathbf{x}^*)\big|\mathbf{x} \sim \mathcal{D}_{id}\big). \tag{4}$$

In general, if the p-value of $\mathbf{x}^*$ is less than 0.05, we can determine that $\mathbf{x}^*$ is an OOD sample at the significance level 0.05.

In practice, **using the p-value is equivalent to using the hard threshold** $S(\mathbf{x}^*) < \lambda$. We denote $\{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^n$ as validation data sampled from the ID distribution $\mathcal{P}_{id}$ and sort their detection score in ascending order: $S_{(1)} \leq S_{(2)} \leq \cdots \leq S_{(n)}$. Since the threshold λ is determined by keeping 95% TPR on the validation data, then we have $S_{([0.05n])} \leq \lambda \leq S_{([0.05n]+1)}$, where $[\cdot]$ is the floor function. On the other hand, the p-value of $\mathbf{x}^*$ less than 0.05 implies that

$$P\big(S(\mathbf{x}) \leq S(\mathbf{x}^*)\big|\mathbf{x} \sim \hat{\mathcal{D}}_{id}\big) \approx 0.05 \;\Rightarrow\; S(\mathbf{x}^*) \lessapprox S_{[0.05n]+1}, \tag{5}$$

where $\hat{\mathcal{D}}_{id}$ is the empirical distribution of $\{\mathbf{x}_i\}_{i=1}^n$. Therefore, when the sample size n is sufficiently large, the OOD region derived from the critical value $\{\mathbf{x} : S(\mathbf{x}) < \lambda\}$ is the same as the OOD region determined by the p-value $\{\mathbf{x} : \text{P-value of } \mathbf{x} < 0.05\}$.

Suppose the test input $\mathbf{x}^*$ is an ID sample that $\mathbf{x}^* \sim \mathcal{D}_{id}$ and the detection score $S(\mathbf{x}^*)$ is a continuous random variable. We write p_0 as the p-value of $\mathbf{x}^*$ and let $F(s)$ be the cumulative distribution function of $S(\mathbf{x})$ with $\mathbf{x} \sim \mathcal{D}_{id}$. Then we have

$$p_0 = P\big(S(\mathbf{x}) \leq S(\mathbf{x}^*)\big|\mathbf{x} \sim \mathcal{D}_{id}\big) = F(S(\mathbf{x}^*)). \tag{6}$$

It follows from the continuity of $S(\mathbf{x}^*)$ and Lemma 21.1 of Van der Vaart (2000) that

$$\begin{aligned} P(p_0 < \alpha) &= 1 - P\big(F(S(\mathbf{x}^*)) \geq \alpha\big) \\ &= 1 - P\big(S(\mathbf{x}^*) \geq F^{-1}(\alpha)\big) = F(F^{-1}(\alpha)) = \alpha. \end{aligned}$$

This implies that **the p-value of $\mathbf{x}^*$ follows a uniform distribution** $U[0, 1]$, if x^* is drawn from the ID distribution. In the following, we will use this property to develop an ensemble scheme.

3.3 TPR CONTROLLING FOR ENSEMBLE

According to Eq. (4), the p-value relies on the score function $S(\mathbf{x})$, which is derived from a pre-trained model ϕ, i.e. $S(x) = S(x; \phi)$. Given one pre-trained model, we can construct a score function and compute the p-value of a test input. But when multiple pre-trained models are accessible, how to fuse the single-model results to leverage the diversity of multiple pre-trained models while strictly maintaining TPR on ID data?

Here we borrow the idea of the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995) and propose an ensemble scheme for OOD detection via p-value correction. Consider a model zoo with m pre-trained models: $\mathcal{M} = \{\phi_1, \phi_2, \ldots, \phi_m\}$ and a score function $S(x; \phi)$. Given a test input $\mathbf{x}^*$ and a pre-trained model ϕ_i, we compute the score value $S(\mathbf{x}^*; \phi_i)$ and obtain the corresponding p-value p_i. Going through all pre-trained models, we obtained m p-values: $\{p_1, p_2, \ldots, p_m\}$, and sort them in ascending order: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. Then, we identify the test input $\mathbf{x}^*$ as an OOD sample if there exists an integer $1 \leq k \leq m$ such that $p_{(k)} \leq \frac{k}{m}(1 - \text{TPR}_0)$. Here '$\text{TPR}_0$' is a predetermined TPR level of the ID data. In general, it is taken to be 95%. We call the proposed method Zoo-based OOD Detection Enhancement (ZODE) and present the details of ZODE in Algorithm 1. Next, we provide theoretical guarantees that Algorithm 1 can maintain the target TPR level on ID data.

Theorem 1 *Suppose a pre-trained model zoo $\{\phi_1, \phi_2, \ldots, \phi_m\}$ is accessible and the score function is $S(x; \phi)$. Let TPR$_0 > 0.5$ be a predetermined TPR level for the ID Data. If the test input $\mathbf{x}^*$ is an ID sample that $\mathbf{x}^* \sim \mathcal{D}_{id}$ and $S(\mathbf{x}^*; \phi_i)$ is independent of $S(\mathbf{x}^*; \phi_j)$ for $\forall i \neq j$, then Algorithm 1 can identify $\mathbf{x}^*$ as an ID data with probability larger than TPR_0.*

Remark. In Theorem 1, we assume that $S(\mathbf{x}^*; \phi_i)$ is independent of $S(\mathbf{x}^*; \phi_j)$, which leads to the independence between p_i and p_j for $\forall i \neq j$. If different pre-trained models learn completely different features, then this assumption can hold. In this case, the model zoo haves the desired diversity. In practice, the pre-trained models can still be very diverse but different models may extract related features. Therefore, we report the empirical TPR of our method in Section 4. One can find that ZODE can still maintain the empirical TPR not less than the target level though the p-values may be related.

Algorithm 1 ZODE: Zoo-based OOD Detection Enhancement

Require: Training data $\{\mathbf{x}_i\}_{i=1}^n$, pre-trained model zoo $\{\phi_1, \ldots, \phi_m\}$, test sample $\mathbf{x}^*$, detection score $S(x; \phi)$, TPR level for ID data 'TPR$_0$';

Compute the score value of $S(\mathbf{x}_i, \phi_j)$, $\forall 1 \leq i \leq n$ and $\forall 1 \leq j \leq m$;
for $1 \leq j \leq m$ **do**
  Compute the empirical distribution of $\{S(\mathbf{x}_i, \phi_j)\}_{i=1}^n$;
  Estimate the p-value of $\mathbf{x}^*$ given ϕ_j:

$$p_j = \frac{\#\{x_i : S(\mathbf{x}_i, \phi_j) \leq S(\mathbf{x}^*, f_j)\}}{m}$$

end for
Sort $\{p_1, \ldots, p_m\}$ in ascending order: $\{p_{(1)}, \ldots, p_{(m)}\}$;
if $\exists 1 \leq j \leq m$ such that $p_{(j)} \leq \frac{j}{m}(1 - \text{TPR}_0)$ **then**
  $\mathbf{x}^*$ is detected as an OOD sample;
else
  $\mathbf{x}^*$ is detected as an ID sample;
end if
return OOD detection decision of $\mathbf{x}^*$.

4 EXPERIMENTS

In this section, we demonstrate the effectiveness of our proposed method. First, we evaluate whether our model zoo and ensemble scheme can enhance OOD detectors. Second, we demonstrate that ZODE exploits the diversity of pre-trained models and leverages the complementarity between the single-model detectors to achieve superior performance. Finally, we show that our method can significantly improve the current SOTA results.

Dataset: We evaluate our proposed method on the CIFAR benchmarks. We use CIFAR10 (Krizhevsky et al., 2009) as the ID data and evaluate OOD detectors on six OOD datasets: SVHN (Netzer et al., 2011), LSUN (Yu et al., 2015), iSUN (Xu et al., 2015), Texture (Cimpoi et al., 2014), Places365 (Zhou et al., 2017), and CIFAR100 (Krizhevsky et al., 2009). We then consider more challenging benchmarks based on ImageNet, i.e., large-scale OOD detection tasks. The ID data is ImageNet-1K (Deng et al., 2009). We evaluate OOD detectors on four test datasets that are subset of : Places365 (Zhou et al., 2017), iNaturalist (Van Horn et al., 2018), SUN (Xiao et al., 2010), and Texture (Cimpoi et al., 2014) with different categories of each other.

Metrics: We evaluate OOD detection methods by the following three metrics: (1) the true positive rate of the ID samples (TPR); (2) the false positive rate of OOD samples when the true positive rate of the ID samples is about 95% (FPR); (3) the area under the receiver operating characteristic curve (AUC). For single-model detectors, the hard threshold is determined by TPR $= 95\%$. Therefore, the first metric aims to check whether our ensemble scheme can maintain the TPR level close to 95%. FPR and AUC are often used in the literature to reflect the capabilities of OOD detectors. For the AUC metric, we use grid values of TPR ranging from 0 to 1 with a gap of 0.0005 and obtain the corresponding FPR to compute the area under the receiver operating characteristic curve.

Table 1: **Results on CIFAR10.** Comparison with competitive OOD detection methods. The results of all competitors are from Sun et al. (2022). All values are percentages. ↓ indicates smaller values are better and vice versa.

Method	OOD Dataset												
		SVHN		**LSUN**		**iSUN**		**Texture**		**Places365**		**Average**	
	TPR	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
MSP	95.00	59.66	91.25	45.21	93.80	54.57	92.12	66.45	88.50	62.46	88.64	57.67	90.86
ODIN	95.00	20.93	95.55	7.26	98.53	33.17	94.65	56.40	86.21	63.04	86.57	36.16	92.30
Energy	95.00	54.41	91.22	10.19	98.05	27.52	95.59	55.23	89.37	42.77	91.02	38.02	93.05
GODIN	95.00	15.51	96.60	4.90	99.07	34.03	94.94	46.91	89.69	62.63	87.31	32.80	93.52
Mahalanobis	95.00	9.24	97.80	67.73	73.61	6.02	98.63	23.21	92.91	83.50	69.56	37.94	86.50
KNN	95.00	24.53	95.69	25.29	95.96	25.55	95.26	27.57	94.71	50.90	89.14	30.77	94.15
CSI	95.00	37.38	94.69	5.88	98.86	10.36	98.01	28.85	94.87	38.31	93.04	24.16	95.89
SSD+	95.00	**1.51**	**99.68**	6.09	98.48	33.60	95.16	12.98	97.70	28.41	94.72	16.52	97.15
KNN+	95.00	2.42	99.52	1.78	99.48	20.06	96.74	8.09	98.56	23.02	95.36	11.07	97.93
ZODE-MSP	95.04	52.44	92.86	15.11	97.62	30.98	95.63	43.16	94.68	43.58	94.55	37.05	95.07
ZODE-Energy	95.07	50.05	92.26	3.12	99.29	16.03	97.09	37.34	95.14	19.52	96.95	25.21	96.15
ZODE-KNN	94.96	2.12	99.43	**1.50**	**99.61**	**5.48**	**98.70**	**0.16**	**99.88**	**9.91**	**97.99**	**3.83**	**99.12**

Table 2: **Results on CIFAR10.** We compare the ZODE-ensembled KNN detector with the single-model KNN detector.

Method	OOD Dataset												
		SVHN		**LSUN**		**iSUN**		**Texture**		**Places365**		**Average**	
	TPR	FPR↓	AUC↑	FPR↓	AUC	FPR↓	AUC↑	FPR↓	AUC	FPR↓	AUC↑	FPR↓	AUC↑
ResNet18	95.00	27.97	95.49	18.50	96.84	24.68	95.52	26.74	94.97	47.95	90.02	29.17	94.57
ResNet18*	95.00	2.42	**99.52**	1.78	99.48	20.06	96.74	8.09	98.57	22.82	95.32	11.03	97.93
ResNet34	95.00	26.53	95.85	10.22	98.39	29.45	95.15	31.65	94.53	36.59	92.75	26.89	95.33
ResNet50	95.00	17.31	97.40	7.10	98.83	17.32	97.26	20.85	96.59	41.35	91.61	20.79	96.34
ResNet101	95.00	25.73	96.12	6.65	98.90	19.84	96.80	18.42	96.89	40.57	92.15	22.24	96.17
ResNet152	95.00	34.96	94.98	7.22	98.88	22.30	96.66	20.76	96.60	38.57	92.36	24.76	95.90
DensNet	95.00	10.22	98.18	7.90	98.60	10.87	97.94	20.78	96.25	50.14	88.92	19.98	95.98
ZODE-KNN	94.96	**2.12**	99.43	**1.50**	**99.61**	**5.48**	**98.70**	**0.16**	**99.88**	**9.91**	**97.99**	**3.83**	**99.12**

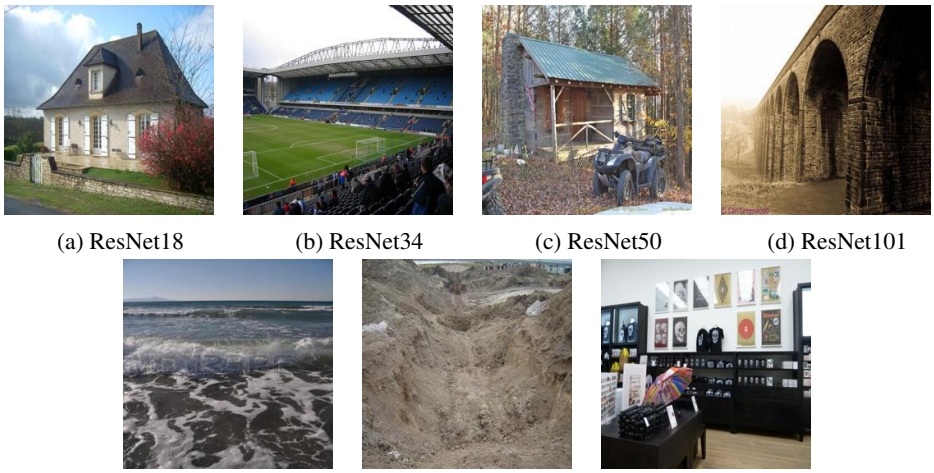
Enhanced OOD detection: We consider three OOD detection methods: MSP (Hendrycks & Gimpel, 2017), Energy (Liu et al., 2020) and KNN (Sun et al., 2022). MSP is a simple baseline method that uses maximum softmax probabilities as the detection score. In some experiments, MSP can yield surprisingly good results when used on top of a large pre-trained model that has been fine-tuned on the ID data (Fort et al., 2021). The energy-based model (LeCun et al., 2006) maps a test input to a scalar that is higher for OOD samples and lower for the training data. Liu et al. (2020) proposes an energy score that uses the logits output by a pre-trained classifier. Sun et al. (2022) uses the feature distance between the test input and the k-th nearest ID sample and proposes a KNN-based detector. These three OOD detection methods represent three kinds of detectors based on probability, logit, and distance, respectively. We take them as the baseline methods and denote our enhanced methods by 'ZODE-MSP', 'ZODE-Energy', and 'ZODE-KNN' respectively.

4.1 Evaluation on CIFAR10 Benchmarks

Model Zoo. We build a model zoo with seven pre-trained models: ResNet18, ResNet34, ResNet50, ResNet101, ResNet152 (He et al., 2016), DenseNet (Huang et al., 2017) and ResNet18* (Sun et al., 2022). Here ResNet and DenseNet are two backbones routinely used in the literature on OOD detection. Therefore, we consider different architectures and use six models trained by cross-entropy loss. In addition, we also notice the effect of the loss function and introduce the model ResNet18* which is trained with contrastive loss. In summary, our model zoo contains diversity derived from different architectures and different training strategies.

ZODE maintains TPR. According to Section 3.1, one of the challenges of ensembling OOD detectors is to control the true positive rate of the ID data. Theorem 1 states that if different pre-trained models learn completely different features, ZODE can keep TPR close to the target level. In Table 1, we report the empirical TPR of ZODE, which is close to the target level 95%.

ZODE-KNN achieves superior performance. We compare our method with competitive OOD detection methods, including MSP (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018), Energy (Liu et al., 2020), GODIN (Hsu et al., 2020), Mahalanobis (Lee et al., 2018), KNN (Sun et al.,



(a) ResNet18 (b) ResNet34 (c) ResNet50 (d) ResNet101

(e) ResNet152 (f) DenseNet (g) ResNet18*

Figure 1: **Places365.** Example OOD images that only one single-model detector can identify.

Table 3: **Results on CIFAR10 for CIFAR100 as OOD.** The results of GRAM and MaSF are from Haroush et al. (2022). We cite the results of SSD and SSD+ reported in (Sehwag et al., 2020).

(a) Comparison with baseline methods

Method	TPR	FPR↓	AUC↑
GRAM	95.00	51.00	83.30
MaSF	95.00	58.20	86.10
SSD	95.00	50.78	90.63
SSD+	95.00	38.50	93.40
KNN	95.00	52.54	89.69
KNN+	95.00	38.83	92.75
ZODE-KNN	94.96	**18.29**	**97.12**

(b) Ensembled vs Single-model

Method	TPR	FPR↓	AUC↑
ResNet18	95.00	52.24	89.69
ResNet18*	95.00	38.83	92.75
ResNet34	95.00	46.74	91.04
ResNet50	95.00	47.14	90.64
ResNet101	95.00	47.07	90.87
ResNet152	95.00	47.72	90.84
DenseNet	95.00	49.43	89.80
ZODE-KNN	94.96	**18.29**	**97.12**

2022), CSI (Tack et al., 2020), SSD+ (Sehwag et al., 2020), as well as KNN+ (Sun et al., 2022). We cite the results of the competitors reported in Sun et al. (2022). For a fair comparison, we set $k = 50$ in the experiments of ZODE-KNN, which is the same as Sun et al. (2022). We can find that compared to the best baseline KNN+, ZODE-KNN reduces the FPR from 11.07% to 3.83%, which significantly improves the relative detection accuracy by 65.40%. Note that ZODE-KNN significantly reduces FPR when OOD samples are drawn from iSUN, Texture, and Places365. For LSUN, ZODE-KNN slightly improves the performance of KNN+. In addition, SSD+ outperforms ZODE-KNN on SVHN. Overall, ZODE-KNN significantly improves the performance of existing methods on these five OOD datasets.

ZODE achieves consistent improvements. We consider three different kinds of OOD detection scores. MSP (Hendrycks & Gimpel, 2017) is based on the probabilities, Energy (Liu et al., 2020) uses the logits, and KNN (Sun et al., 2022) directly quantifies the distance in the embedding space. Then we compare them with the corresponding enhanced detectors: ZODE-MSP, ZODE-Energy, and ZODE-KNN. For ZODE-MSP and ZODE-Energy, we use the same settings as Hendrycks & Gimpel (2017) and Liu et al. (2020) respectively. We find that ZODE-enhanced detectors consistently improve the performance of the corresponding baselines.

ZODE leverages the complementarity between the single-model detectors. Table 2 reports the results of all single-model detectors derived from our model zoo and KNN score. It is easy to see that the ZODE-ensembled KNN detector significantly outperforms all single-model KNN detectors on LSUN, iSUN, Texture, and Places365. Compared with the best single-model baseline, ZODE reduces the FPR from 11.03% to 3.83%, which significantly improves the relative detection accuracy by 65.28%. Moreover, ZODE improves the performance sharply on Texture and Places365. This

Table 4: **Results on ImageNet.** All results of the competitors are cited from Sun et al. (2022). Methods reported are all based on ID data only (ImageNet-1k).

Method	OOD Dataset										
		iNaturalist		**SUN**		**Places**		**Textures**		**Average**	
	TPR	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
MSP	95.00	54.99	87.74	70.83	80.86	73.99	79.76	68.00	79.61	66.95	81.99
ODIN	95.00	47.66	89.66	60.15	84.59	67.89	81.78	50.23	85.62	56.48	85.41
Energy	95.00	55.72	89.95	59.26	85.89	64.92	82.86	53.72	85.99	58.41	86.17
GODIN	95.00	61.91	85.40	60.83	85.60	63.70	83.81	77.85	73.27	66.07	82.02
Mahalanobis	95.00	97.00	52.65	98.50	42.41	98.40	41.79	55.80	85.01	87.43	55.47
KNN	95.00	59.00	86.47	68.82	80.72	76.28	75.76	11.77	97.07	53.97	85.01
SSD+	95.00	57.16	87.77	78.23	73.10	81.19	70.97	36.37	88.52	63.24	80.09
KNN+	95.00	30.18	94.89	48.99	88.63	59.15	84.71	15.55	95.40	38.47	90.91
ZODE-KNN	94.89	**5.01**	**98.60**	**48.87**	**90.37**	**53.96**	**88.07**	**4.57**	**98.93**	**28.10**	**93.99**

implies that the superior performance of ZODE does not fully come from any single-model detector. Therefore, our ensemble procedure works and is necessary for the improvements.

We further take Place365 as an example to illustrate that ZODE exploits the diversity of multiple pre-trained models. At step 7 of Algorithm 1, if $p_{(1)} \leq \frac{1}{m}(1 - \text{TPR}_0)$ and $p_{(j)} > \frac{j}{m}(1 - \text{TPR}_0)$, $\forall j \geq 2$, then there is only one pre-trained model that can help to identify the test input as an OOD sample. Figure 1 presents seven such images and each image corresponds to one pre-trained model in our model zoo.

Evaluations on CIFAR10 vs CIFAR100. We consider a challenging OOD detection task that identifies OOD samples drawn from CIFAR100 when the ID data is CIFAR10. Table 3a summarizes a detailed comparison with GRAM (Sastry & Oore, 2019), MaSF (Haroush et al., 2022), SSD (Sehwag et al., 2020), and KNN (Sun et al., 2022). Compared with the best baseline SSD+, ZODE reduces the FPR by 20.21%, which is a relative 52.49% improvement in detection power. The results in Table 3b clearly show that ZODE significantly outperforms the single-model-based KNN detectors and our ensemble scheme fully leverages the complementarity between the single-model detectors.

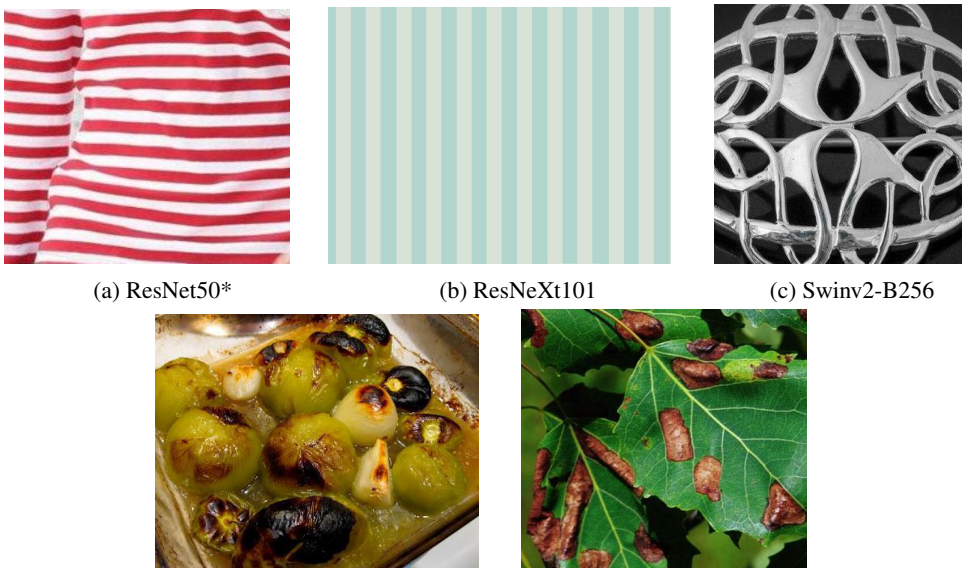
4.2 Evaluation on ImageNet Benchmarks

Model zoo and implementation details. We use five pre-trained models to build a model zoo, consisting of models with different architectures and different pre-training strategies. The models are as follows: ResNet50* (Sun et al., 2022), semi-weekly supervised ResNeXt101 32x16d (Yalniz et al., 2019), Swinv2-B256, Swinv2-B384, and Swinv2-L256 (Liu et al., 2022). Significantly, resolutions of Swinv2-B256, Swinv2-B384, and Swinv2-L256 are 256x256, 256x256, and 384x384 respectively. ResNet50* is trained with SupCon loss (Khosla et al., 2020), which pulls points belonging to the same class together in the embedding space and separates samples from different classes. ResNeXt101 is pre-trained on Billion-scale images associated with meta information semantically relevant to ImageNet, which achieves 84.8% top-1 accuracy on ImageNet. The three Swinv2 models are pre-trained at higher resolution, and their top-1 accuracy on Imagenet all exceed 84%. In the following, we only report the results of ZODE-KNN based on the model zoo. The hyperparameter TPR_0 is taken to be 93.50%, which makes the empirical TPR of ZODE-KNN close to 95%. We use $k = 1000$ for ResNet50*, which is same as Sun et al. (2022). For the rest models, we selected k from $\{100, 200, 500, 700, 800, 900, 1000, 3000, 5000\}$ that minimize the FPR.

ZODE+KNN achieves superior performance. In Table 4, we compare ZODE-KNN with competitive OOD detection methods, including MSP (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018), Energy (Liu et al., 2020), GODIN (Hsu et al., 2020), Mahalanobis (Lee et al., 2018), KNN (Sun et al., 2022), SSD+ (Sehwag et al., 2020), as well as KNN+ (Sun et al., 2022). ZODE-KNN outperforms the best baseline KNN+ uniformly on all four OOD datasets, substantially reducing the average FPR from 38.47% to 28.10%, which achieves a relative 26.96% improvement in detection power. Especially when test datasets are iNaturalist and Textures, ZODE-KNN reduces the relative FPR by 83.40% and 70.61% respectively, which highlights the effectiveness of ZODE.

Table 5: **Results on ImageNet.** Comparison with single-model detectors and ZODE.

Method	OOD Dataset										
		iNaturalist		**SUN**		**Places**		**Textures**		**Average**	
	TPR	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
ResNet50*	95.00	30.18	94.89	48.99	88.63	59.15	84.71	15.55	95.40	38.47	90.91
ResNext101 32x16	95.00	15.24	96.78	56.06	88.60	61.74	86.29	26.06	93.53	39.78	91.30
Swinv2-B256	95.00	9.11	97.93	58.16	88.78	58.66	87.13	41.24	89.67	41.79	90.88
Swinv2-B384	95.00	5.65	98.50	49.59	90.28	**52.27**	**88.44**	38.37	89.99	36.47	91.80
Swinv2-L256	95.00	6.98	98.44	52.43	89.49	53.81	88.07	39.26	89.92	38.12	91.48
ZODE-KNN	94.89	**5.01**	**98.60**	**48.87**	**90.37**	53.96	88.07	**4.57**	**98.93**	**28.10**	**93.99**



(a) ResNet50* (b) ResNeXt101 (c) Swinv2-B256

(d) Swinv2-B384 (e) Swinv2-L256

Figure 2: Textures. Example OOD images that only one single-model detector can identify.

ZODE combines the advantages of the single-model detectors. In Table 5, we report the performance of every single-model detector derived from our model zoo. We highlight three trends: (1) ZODE-KNN outperforms the best single-model KNN detector with a relative 22.95% improvement in FPR. This implies that ZODE works in the ImageNet benchmarks and the ensemble scheme of ZODE-KNN is necessary for the improvements. (2) ZODE combines the advantages of single-model detectors. In Table 5, we can observe that ResNet50* and ResNeXt101 32x16 perform well on Textures, but underperform on iNaturalist, while the Swin models show the opposite performance. However, the ZODE-ensembled detector achieves strong and stable performance in all test datasets. (3) ZODE leverages the complementarity between the single-model detectors. Similar to the discussions in Figure 1, we find some images in Textures that can be successfully identified as OOD samples and the detection decision depends only on one single-model detector. Figure 1 presents five such images and each image corresponds to one pre-trained model in our model zoo.

5 Conclusion

In this paper, we exploit the diversity of multiple pre-trained models in a model zoo to improve the performance of post hoc OOD detection. We propose, ZODE, an efficient and fundamental ensemble scheme for combining multiple detection decisions. Extensive experiments show that ZODE can effectively solve the missed detection problem of single-model detectors by exploiting the complementarity of multiple detectors. We find that ZODE combined with the KNN detector Sun et al. (2022) works very well. On a wide range of OOD detection benchmarks, ZODE-KNN significantly improves the current SOTA results.

REFERENCES

Felix Abramovich and Ya'acov Ritov. *Statistical theory: a concise introduction*. CRC Press, 2013.

Jerone Andrews, Thomas Tanay, Edward J Morton, and Lewis D Griffin. Transfer representation-learning for anomaly detection. JMLR, 2016.

Abhijit Bendale and Terrance Boult. Towards open world recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1893–1902, 2015.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pp. 1165–1188, 2001.

Federico Bergamin, Pierre-Alexandre Mattei, Jakob Drachmann Havtorn, Hugo Senetaire, Hugo Schmutz, Lars Maaløe, Soren Hauberg, and Jes Frellsen. Model-agnostic out-of-distribution detection using combined statistical tests. In *International Conference on Artificial Intelligence and Statistics*, pp. 10753–10776. PMLR, 2022.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Raghavendra Chalapathy and Sanjay Chawla. Deep learning for anomaly detection: A survey. *arXiv preprint arXiv:1901.03407*, 2019.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

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. URL `https://openreview.net/forum?id=YicbFdNTTy`.

Linus Ericsson, Henry Gouk, and Timothy M Hospedales. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5414–5423, 2021.

Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Matan Haroush, Tzviel Frostig, Ruth Heller, and Daniel Soudry. A statistical framework for efficient out of distribution detection in deep neural networks. In *International Conference on Learning Representations*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

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

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10951–10960, 2020.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

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

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Alex Krizhevsky, Geoffrey Hinton, 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. *Advances in neural information processing systems*, 30, 2017.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A tutorial on energy-based learning, 2006.

Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15313–15323, 2021.

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

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Akshayaa Magesh, Venugopal V Veeravalli, Anirban Roy, and Susmit Jha. Multiple testing framework for out-of-distribution detection. *arXiv preprint arXiv:2206.09522*, 2022.

David J Miller, Zhen Xiang, and George Kesidis. Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks. *Proceedings of the IEEE*, 108 (3):402–433, 2020.

Warren Morningstar, Cusuh Ham, Andrew Gallagher, Balaji Lakshminarayanan, Alex Alemi, and Joshua Dillon. Density of states estimation for out of distribution detection. In *International Conference on Artificial Intelligence and Statistics*, pp. 3232–3240. PMLR, 2021.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 427–436, 2015.

Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441: 138–150, 2021.

Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, Zahra Moayed, and Reinhard Klette. Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes. *Computer Vision and Image Understanding*, 172:88–97, 2018.

Chandramouli S Sastry and Sageev Oore. Zero-shot out-of-distribution detection with feature correlations. 2019.

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 *International conference on information processing in medical imaging*, pp. 146–157. Springer, 2017.

Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2020.

Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 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*, pp. 20827–20840. PMLR, 17–23 Jul 2022.

Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on distributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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.

Volodya Vovk, Alexander Gammerman, and Craig Saunders. Machine-learning applications of algorithmic randomness. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 444–453, 1999.

Apoorv Vyas, Nataraj Jammalamadaka, Xia Zhu, Dipankar Das, Bharat Kaul, and Theodore L Willke. Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 550–564, 2018.

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, 2022.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, et al. Contrastive training for improved out-of-distribution detection. *arXiv preprint arXiv:2007.05566*, 2020.

Jianxiong Xiao, J Hays, KA Ehinger, A Oliva, and A Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

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.

I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546*, 2019.

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.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

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.

A Proof of Theorem 1

Theorem. *Suppose a pre-trained model zoo $\{\phi_1, \phi_2, \ldots, \phi_m\}$ is accessible and the score function is $S(x;\phi)$. Let TPR$_0$ > 0.5 be the target TPR level for the ID data. If the test input $\mathbf{x}^*$ is an ID sample that $\mathbf{x}^* \sim \mathcal{D}_{id}$ and $S(\mathbf{x}^*;\phi_i)$ is independent of $S(\mathbf{x}^*;\phi_j)$ for $\forall i \neq j$, then Algorithm 1 can identify $\mathbf{x}^*$ as an ID data with probability larger than TPR_0.*

Proof. According to Section 3.2, $p_1, ..., p_m$ follows uniform distribution, with density function

$$f_{p_i}(x) = \begin{cases} 1 & x \in [0,1]; \\ 0 & otherwise. \end{cases}$$

Then, the joint probability density of the ordered values $p_{(1)}, p_{(2)}, ..., p_{(m)}$ is

$$f_{p_{(1)},p_{(2)},...,p_{(m)}}(x_1, x_2, ..., x_m) = m! \prod_{i=1}^{S} f_{p_i}(x_i) = m!$$

We denote $\alpha = 1 - \text{TPR}_0$ and define an event $\mathcal{E}$: $\forall\, 1 \leq j \leq m,\ p_{(j)} \geq \frac{j}{m}\alpha$. Then we have

$$\begin{aligned} P(\mathcal{E}|\mathbf{x}^* \sim \mathcal{D}_{id}) &= \int_{\frac{m}{m}\alpha}^{1} ... \int_{\frac{2}{m}\alpha}^{1} \int_{\frac{1}{m}\alpha}^{1} f_{p_{(1)},\cdots,p_{(m)}}(x_1, x_2, \cdots, x_m) \mathrm{d}x_1 \mathrm{d}x_2 \cdots \mathrm{d}x_m \\ &= m!(1 - \frac{1}{m}\alpha)(1 - \frac{2}{m}\alpha)...(1 - \frac{m}{m}\alpha). \end{aligned} \tag{7}$$

Next, we prove that for any $m \geq 1$ and $\alpha \leq 0.5$,

$$m!(1 - \frac{1}{m}\alpha)(1 - \frac{2}{m}\alpha)...(1 - \frac{m}{m}\alpha) \geq 1 - \alpha \tag{8}$$

It is easy to see that Eq. (8) holds when $m = 1$. Suppose Eq. (8) holds for $m = m_0$. Then for $m = m_0 + 1$, we have

$$\begin{aligned} &(m_0+1)!(1 - \frac{1}{m_0+1}\alpha)(1 - \frac{2}{m_0+1}\alpha)...(1 - \frac{m_0+1}{m_0+1}\alpha) \\ \geq &(m_0+1)!(1 - \frac{1}{m_0}\alpha)(1 - \frac{2}{m_0}\alpha)...(1 - \frac{m_0}{m_0}\alpha)(1 - \frac{m_0+1}{m_0+1}\alpha) \\ \geq &(m_0+1)(1-\alpha)^2 \geq 1 - \alpha, \end{aligned} \tag{9}$$

which implies that Eq.(8) also holds for $m = m_0 + 1$. Hence, the proof is finished.

□

Table 6: **Results on CIFAR10.** Comparison with competitive OOD detection methods. The results of all competitors are from Sun et al. (2022). All values are percentages. ↓ indicates smaller values are better and vice versa.

Method	OOD Dataset													
		SVHN		**LSUN**		**iSUN**		**Texture**		**Places365**		**Average**	
	TPR	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
MSP	95.00	59.66	91.25	45.21	93.80	54.57	92.12	66.45	88.50	62.46	88.64	57.67	90.86
ODIN	95.00	20.93	95.55	7.26	98.53	33.17	94.65	56.40	86.21	63.04	86.57	36.16	92.30
Energy	95.00	54.41	91.22	10.19	98.05	27.52	95.59	55.23	89.37	42.77	91.02	38.02	93.05
GODIN	95.00	15.51	96.60	4.90	99.07	34.03	94.94	46.91	89.69	62.63	87.31	32.80	93.52
Mahalanobis	95.00	9.24	97.80	67.73	73.61	6.02	98.63	23.21	92.91	83.50	69.56	37.94	86.50
KNN	95.00	24.53	95.69	25.29	95.96	25.55	95.26	27.57	94.71	50.90	89.14	30.77	94.15
CSI	95.00	37.38	94.69	5.88	98.86	10.36	98.01	28.85	94.87	38.31	93.04	24.16	95.89
SSD+	95.00	**1.51**	**99.68**	6.09	98.48	33.60	95.16	12.98	97.70	28.41	94.72	16.52	97.15
KNN+	95.00	2.42	99.52	1.78	99.48	20.06	96.74	8.09	98.56	23.02	95.36	11.07	97.93
ZODE-KNN(k=1)	95.00	2.60	99.42	2.34	99.46	7.29	98.47	0.44	99.73	11.79	97.77	4.89	98.97
ZODE-KNN(k=50)	94.96	2.12	99.43	**1.50**	**99.61**	**5.48**	**98.70**	**0.16**	**99.88**	**9.91**	**97.99**	**3.83**	**99.12**

Table 7: **Results on CIFAR10**. Comparison with single-model detectors at different k levels. Setting $k = 1$ and $k = 50$ respectively, we compare the performance with single-model detectors and ZODE.

K	Method	OOD Dataset															
			SVHN		**LSUN**		**iSUN**		**Texture**		**Places365**		**CIFAR100**		**Average**	
		TPR	FPR↓	AUC↑	FPR↓	AUC	FPR↓	AUC↑	FPR↓	AUC	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
$k = 1$	ResNet18	95.00	50.34	91.45	25.44	95.67	30.53	94.71	33.44	93.69	48.40	90.05	53.30	89.44	40.24	92.50
	ResNet18*	95.00	**2.16**	**99.56**	4.45	98.78	21.99	96.56	9.88	98.34	23.87	95.38	39.88	92.90	17.04	96.92
	ResNet34	95.00	28.09	95.55	11.80	98.13	30.90	94.97	32.19	94.24	38.67	92.41	47.92	90.82	31.71	94.35
	ResNet50	95.00	22.37	96.25	7.15	98.67	18.40	96.90	23.63	95.72	41.00	91.64	46.81	90.45	26.56	94.94
	ResNet101	95.00	25.45	95.96	8.46	98.62	20.92	96.54	21.03	96.31	41.07	92.18	47.71	90.77	27.44	95.07
	ResNet152	95.00	33.79	94.16	8.48	98.63	19.06	96.93	20.62	96.50	37.97	92.52	46.77	91.01	27.78	94.96
	DensNet	95.00	13.46	97.76	9.41	98.31	13.70	97.58	24.29	95.46	51.49	89.07	50.01	89.80	27.06	94.66
	ZODE-KNN	95.00	2.60	99.42	**2.34**	**99.46**	**7.29**	**98.47**	**0.44**	**99.73**	**11.79**	**97.77**	**20.58**	**96.85**	**7.51**	**98.62**
$k = 50$	ResNet18	95.00	27.97	95.49	18.50	96.84	24.68	95.52	26.74	94.97	47.95	90.02	52.24	89.69	33.01	93.75
	ResNet18*	95.00	2.42	**99.52**	1.78	99.48	20.06	96.74	8.09	98.57	22.82	95.32	38.83	92.75	15.67	97.06
	ResNet34	95.00	26.53	95.85	10.22	98.39	29.45	95.15	31.65	94.53	36.59	92.75	46.74	91.04	30.20	94.62
	ResNet50	95.00	17.31	97.40	7.10	98.83	17.32	97.26	20.85	96.59	41.35	91.61	47.14	90.64	25.18	95.39
	ResNet101	95.00	25.73	96.12	6.65	98.90	19.84	96.80	18.42	96.89	40.57	92.15	47.07	90.87	26.38	95.29
	ResNet152	95.00	34.96	94.98	7.22	98.88	22.30	96.66	20.76	96.60	38.57	92.36	47.72	90.84	28.59	95.06
	DensNet	95.00	10.22	98.18	7.90	98.60	10.87	97.94	20.78	96.25	50.14	88.92	49.43	89.80	24.89	94.95
	ZODE-KNN	94.96	**2.12**	99.43	**1.50**	**99.61**	**5.48**	**98.70**	**0.16**	**99.88**	**9.91**	**97.99**	**18.29**	**97.12**	**6.24**	**98.79**

B INFLUENCE OF HYPERPARAMETER

In Secction 4, we find that the ZODE-enhanced KNN detector works very well and significantly improves the best baseline method on a wide range of OOD detection benchmarks. The KNN detector Sun et al. (2022) uses the distance between the test input and the k-th nearest ID sample as the detection score. Here k is a hyperparameter that needs to be predetermined. In this section, we use CIFAR10 to investigate the effect of the choice of k on ZODE. We consider $k = 1$ and $k = 50$, and report the results in Table 6. One can find that the choice of k affects the performance of ZODE-KNN. But the influence is not significant compared to the improvements of ZODE-KNN. Both ZODE-KNN(k=1) and ZODE-KNN(k=50) significantly outperform the best baseline methods. In Table 7, we report the detailed comparison between the ZODE-ensembled KNN detector and the single-model KNN detectors derived from our model zoo. We observe a similar phenomenon in that the choice of k affects the performance of both ensembled and single-model detectors, while the influence is not significant compared to the improvement caused by the ensemble scheme.

C MORE RELATED WORK

Score function for OOD detection. Many recent works have proposed to use models pre-trained with multi-class classification tasks to do OOD detection. A natural idea to distinguish OOD samples is to use the softmax confidence score based on the classifier output and a normalized probability vector. However, neural networks might give an OOD sample a high confidence prediction (Guo et al., 2017; Hein et al., 2019). To deal with this issue, several studies have improved scoring methods and also proposed new scoring measures, such as OpenMax score (Bendale & Boult, 2015), maximum softmax probability (Hendrycks & Gimpel, 2017), ODIN score (Liang et al., 2018), deep ensembles (Lakshminarayanan et al., 2017), Mahalanobis distance-based score (Lee et al., 2018), Gram matrix (Sastry & Oore, 2019), energy score (Liu et al., 2020), (Lin et al., 2021), activation

rectification (ReAct) (Sun et al., 2021), gradient-based score (GradNorm) (Huang et al., 2021), and ViM score (Wang et al., 2022). In this work, we consider KNN score Sun et al. (2022) that employ the non-parametric nearest neighbor approach to quantify the distance between the test input and the training data.

Model-training strategies for OOD detection. Besides proposing efficient score functions, another area of research focuses on choosing strategies for training deep models, including generative adversarial networks (Schlegl et al., 2017), convolutional network (Sabokrou et al., 2018) and transfer representation-learning (Andrews et al., 2016). These classical neural network architectures, such as ResNet (He et al., 2016), DenseNet (Huang et al., 2017) and Swin (Liu et al., 2021) etc, show different effects in downstream tasks when trained with different training strategies or regularization methods (Ericsson et al., 2021). For example, utilizing auxiliary OOD training data like artificially synthesized data from GANs (Lee et al., 2017) and unlabeled data for outlier exposure (Hendrycks et al., 2018) can explicitly regularize the model. Another method to fine-tune the model is to modify the loss function or use auxiliary objectives. Some loss functions encourage the predictive distribution of OOD sample toward uniform distribution (Lee et al., 2017; Hendrycks et al., 2018), and also some by adding contrastive loss (Winkens et al., 2020), margin loss (Vyas et al., 2018) or objective function of adversarial learning (Biggio & Roli, 2018; Miller et al., 2020; Chalapathy & Chawla, 2019) can force models to learn more high-level, task-agnostic, comprehensive features from the training dataset, to enable it robust enough for downstream tasks with various distribution shifts. These works either modify the neural network and objective function or require additional data, at the cost of computational cost.