Enhancing the Power of OOD Detection via Sample-Aware Model Selection

Feng Xue*
Shanghai Jiao Tong University
sjtu2988237@sjtu.edu.cn

Zi He*
Hunan University
hezi0107@hnu.edu.cn

Yuan Zhang*
Beijing Normal University
1256033547@qq.com

Chuanlong Xie
Beijing Normal University
clxie@bnu.edu.cn

Zhenguo Li
Huawei Noah's Ark Lab
li.zhenguo@huawei.com

Falong Tan†
Hunan University
falongtan@hnu.edu.cn

Abstract

In this work, we present a novel perspective on detecting out-of-distribution (OOD) samples and propose an algorithm for sample-aware model selection to enhance the effectiveness of OOD detection. Our algorithm determines, for each test input, which pre-trained models in the model zoo are capable of identifying the test input as an OOD sample. If no such models exist in the model zoo, the test input is classified as an in-distribution (ID) sample. We theoretically demonstrate that our method maintains the true positive rate of ID samples and accurately identifies OOD samples with high probability when there are a sufficient number of diverse pre-trained models in the model zoo. Extensive experiments were conducted to validate our method, demonstrating that it leverages the complementarity among single-model detectors to consistently improve the effectiveness of OOD sample identification. Compared to baseline methods, our approach improved the relative performance by 65.40% and 37.25% on the CIFAR10 and ImageNet benchmarks, respectively.

1. Introduction

Deep neural networks have shown remarkable success in a variety of applications, but their ability to generalize robustly remains a challenging issue in deep learning. While highly trained and complex deep neural networks can perform exceptionally well on test data that is identically distributed (ID) with the training data, their effectiveness in accurately predicting inputs that fall outside of the training distribution is limited. This poses a significant hurdle to the generalization capability of deep neural network models. In safety-critical applications, it is preferable to detect out-of-distribution (OOD) inputs beforehand rather than relying on the model to make potentially unreliable predictions.

Utilizing pre-trained network models, post hoc OOD detection has shown significant potential in addressing large-scale problems. The post hoc detection method typically involves two crucial steps: (i) selecting a pre-trained model that captures the distinction between OOD samples and the in-distribution (ID) samples, and (ii) generating a test score that measures the similarity of a given test input to the ID samples used for training. Various score functions have been developed to differentiate OOD samples by utilizing different outputs of pre-trained models [11, 18], energies [19, 32], and features [8, 16, 17, 24, 26–29, 36].Considerable research effort has been dedicated to this field.

Despite significant advancements in deep neural network architecture, the selection of an appropriate pre-trained model remains a challenging initial step. This is because the effectiveness of post hoc detection heavily relies on the choice of pre-trained models [7]. In Section 5, Table 2 presents the performance of various single models using the KNN detector [28] for OOD detection on ID samples from CIFAR10. Notably, when the OOD samples are sourced from SVHN [22], there is a significant difference of approximately 32.54% in the false positive rate (FPR) between the best-performing and worst-performing models. Furthermore, during the training process, both OOD samples and OOD distributions are unknown.[1] Consequently, hyperparameter tuning and model selection solely rely on the ID or training data. Once deployed in an open-world setting, the OOD detector may encounter a diverse range of test inputs originating from various OOD distributions. Therefore, it is possible that a single pre-trained model may not be capable of handling potential distribution shifts (See Figure 1 in Section 5). Hence, adaptive model selection is crucial and should be performed on a per-test-sample basis.

[*] Equal contribution.
[†] Corresponding author.

[1] In this work, we consider OOD detection without exposure to auxiliary data or OOD samples.

However, there is currently no clear guidance on how to select a pre-trained model in advance for an OOD detection task. As a result, most existing methods rely on trial-and-error or empirical heuristics. In this study, we utilize a collection of pre-trained models (i.e., model zoo) to address the challenges associated with selecting a suitable model for OOD detection. Firstly, we establish a model zoo that includes various network structures and pre-training strategies. This allows us to capture a wide range of input properties and effectively handle different distribution changes. Next, we redefine OOD detection by evaluating whether any model in the model zoo can identify a test input as an OOD sample. Essentially, this can be viewed as a sample-aware model selection task, where the objective is to identify a pre-trained model from the model zoo that can effectively distinguish a given test input from the in-distribution (ID) data. If there is no pre-trained model in the model zoo capable of making this distinction, the test input is classified as an ID sample.

In this study, we introduce a novel out-of-distribution (OOD) detector called ZODE, which stands for Zoo-based OOD Detection Enhancement. Our approach incorporates sample-aware model selection and integrates multiple OOD detection decisions obtained from a model zoo. To achieve this, we utilize p-values for normalization and subsequently adjust their threshold values. The most effective pre-trained models are selected using the Benjamini-Hochberg correction method [2]. Theoretical analysis demonstrates that ZODE can maintain a high true positive rate (TPR) while achieving a low false positive rate (FPR) on OOD samples. By comparing the performance of single-model detectors with that of ZODE, we observe that our proposed method effectively harnesses the diversity and complementarity of multiple pre-trained models.

Our contributions can be summarized as follows:

- We offer a new perspective on OOD detection by incorporating the concept of a model zoo.Additionally, we propose ZODE, a sample-aware model selection algorithm, for OOD detection.
- Through a thorough theoretical analysis of ZODE, we demonstrate its effectiveness in maintaining a high true positive rate (TPR) while achieving a low false positive rate (FPR). Furthermore, our approach leverages the complementarity among single-model detectors to enhance overall performance.
- To validate the efficacy and consistency of our method, we conduct extensive experiments on both the CIFAR10 benchmark and a challenging OOD detection task based on ImageNet. Specifically, our method yields a significant improvement in the average FPR from 11.07% to 3.83% for CIFAR10. Similarly, for the ImageNet task, our method reduces the average FPR from 38.47% to 24.14%.

2. Preliminaries

Out-of-Distribution (OOD) Detection is a task that aims to determine whether a test input is generated from the training distribution. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and label spaces, respectively, and P_0 denote the training distribution over $\mathcal{X} \times \mathcal{Y}$. We denote the marginal distribution of P_0 on $\mathcal{X}$ as $P_{\mathbf{x}}$. An input sample $\mathbf{x} \sim P_{\mathbf{x}}$ is referred to as an in-distribution (ID) sample, while an "unknown" input is identified as an OOD data. Generally, the OOD detection task can be formulated as a hypothesis-testing problem:

$$\mathcal{H}_0 : \mathbf{x}^* \sim P_{\mathbf{x}} \quad \text{versus} \quad \mathcal{H}_1 : \mathbf{x}^* \sim Q \in \mathcal{Q}, \tag{1}$$

where $\mathcal{Q}$ is a set of distributions and $P_{\mathbf{x}} \notin \mathcal{Q}$. The problem (1) does not assume a specific OOD distribution in the alternative hypothesis $\mathcal{H}_1$. Furthermore, we only have access to the ID data sampled from P_0, making it a typical one-sample hypothesis-testing problem.

Consider a pre-trained neural network $\phi(\mathbf{x})$ and an OOD detector that distinguishes ID and OOD samples at test time using a decision function:

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

Here, $\mathbf{x}^*$ is a test input, $S(\cdot;\phi)$ is a score function that assigns higher scores to ID data and lower scores to OOD data, and λ_ϕ is a threshold value. We denote $F(s;\phi)$ as the distribution of $S(\mathbf{x};\phi)$ with $\mathbf{x} \sim P_{\mathbf{x}}$. To maintain a true positive rate (TPR) of ID samples at probability $1-\alpha$, we choose λ_ϕ as the α-quantile of $F(s;\phi)$. Let $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ be the validation set of ID data. Then, the empirical distribution of $S(\mathbf{x};\phi)$ is given by

$$\hat{F}(s;\phi) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\big\{S(\mathbf{x}_i;\phi) \le s\big\}, \tag{3}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function. The threshold λ_ϕ is computed as $\lambda_\phi = \hat{F}^{-1}(\alpha;\phi) = \inf_{s\in\mathbb{R}}\{s : \hat{F}(s;\phi) \ge \alpha\}$.

3. Methodology

3.1. Challenges in Model Selection

Given a collection of pre-trained neural network models $\mathcal{M} = \{\phi_1, \ldots, \phi_m\}$, a common approach for model selection is to employ Eq.(2) to evaluate each pre-trained model and then determine the set of active models that can detect an out-of-distribution (OOD) sample $\mathbf{x}^*$, denoted by $\mathcal{A}(\mathbf{x}^*;\mathcal{M}) = \big\{\phi : \phi \in \mathcal{M}, G(\mathbf{x}^*;\phi) = \text{OOD}\big\}$. Using this set, we can construct a naive detector as follows:

$$G(\mathbf{x}^*;\mathcal{M}) = \begin{cases} \text{ID} & \text{if} \quad \mathcal{A}(\mathbf{x}^*;\mathcal{M}) = \emptyset; \\ \text{OOD} & \text{if} \quad \mathcal{A}(\mathbf{x}^*;\mathcal{M}) \ne \emptyset. \end{cases} \tag{4}$$

In other words, a test input $\mathbf{x}^*$ is classified as an in-distribution (ID) sample only if all detectors $G(\mathbf{x}^*; \phi_i)$, where $\phi_i \in \mathcal{M}$, agree that $\mathbf{x}^*$ is ID. This approach can be improved by leveraging additional information such as the confidence level of each detector. However, this simple approach is unreliable as it fails to maintain the true positive rate (TPR) of the ID data. Let us denote the target TPR level as $1 - \alpha$. Each detector $G(\mathbf{x}^*; \phi_i)$ has a probability α of incorrectly identifying an ID sample as an OOD sample. When combining these detection decisions, the probability of error accumulation increases. It is evident that when the detectors are independent, this naive ensembled detector $G(\mathbf{x}^*; \mathcal{M})$ can misclassify an ID sample as an OOD sample with a probability of $1 - (1 - \alpha)^m$. As the number of pre-trained models (i.e., m) in the model zoo increases ($m \to +\infty$), the error probability also increases, eventually reaching 100%. This indicates that the naive ensembled detector $G(\mathbf{x}^*; \mathcal{M})$ cannot maintain the target TPR level. In this work, we propose an adjustment scheme for threshold values that can maintain the TPR while simultaneously achieving a high probability of correctly identifying OOD data (low false positive rate).

3.2. Normalizing Detection Decisions

The detection score $G(\mathbf{x}; \phi)$ can exhibit variations in range, scale, and distribution across different pre-trained models. To unify and normalize multiple OOD detection decisions, we employ the p-value [1]. The p-value is a probability measure that quantifies the degree of extremity of the observed score if the test input comes from the ID distribution. Given a test sample $\mathbf{x}^*$ and its detection score $\mathbf{s}^* = S(\mathbf{x}^*; \phi)$, the p-value of $\mathbf{x}^*$ is defined as

$$p = \mathbb{P}\big(S(\mathbf{x}; \phi) \leq \mathbf{s}^* \big| \mathbf{x} \sim P_{\mathbf{x}}\big) = F(\mathbf{s}^*; \phi), \tag{5}$$

where $F(\cdot)$ is the cumulative distribution function of the detection score for an ID sample. If the p-value of $\mathbf{x}^*$ is smaller than a significance level α, then $\mathbf{x}^*$ is classified as an OOD sample. It is apparent that using the p-value is equivalent to employing a hard threshold, denoted as λ in Eq. (2).

Suppose that $\mathbf{x}^*$ is an ID sample, where $\mathbf{x}^* \sim P_{\mathbf{x}}$, and the detection score $\mathbf{s}^* = S(\mathbf{x}^*; \phi)$ is a continuous random variable. According to the continuity of $\mathbf{s}^*$ and Lemma 21.1 of Van der Vaart [30], we have

$$\begin{aligned}\mathbb{P}(p < \alpha) &= 1 - \mathbb{P}\big(F(\mathbf{s}^*; \phi) \geq \alpha\big)\\ &= 1 - \mathbb{P}\big(\mathbf{s}^* \geq F^{-1}(\alpha)\big)\\ &= F(F^{-1}(\alpha)) = \alpha,\end{aligned}$$

where p is the p-value of $\mathbf{x}^*$, and $F^{-1}(\alpha)$ is the inverse cumulative distribution function of the detection score for an ID sample. This implies that the p-values of an ID sample follow a uniform distribution $U[0, 1]$, and this result holds for any $\phi \in \mathcal{M}$. In the following, we leverage this property to develop the threshold adjustment.

Algorithm 1 ZODE: Zoo-based OOD Detection Enhancement

```
Require: Test sample $\mathbf{x}^*$, validation set of training data $\{\mathbf{x}_i\}_{i=1}^n$, pre-trained model zoo $\mathcal{M} = \{\phi_1, \ldots, \phi_m\}$, score function $S(x; \phi)$, target TPR level $1 - \alpha$;
for $1 \leq j \leq m$ do
    Compute $\mathbf{s}_j^* = S(\mathbf{x}^*; \phi_j)$;
    Compute $p_j = \hat{F}(\mathbf{s}_j^*; \phi_j)$ according to Eq. (3);
end for
Sort $\{p_1, \ldots, p_m\}$ in ascending order and denote the ranked $p$-values by $\{p_{(1)}, \ldots, p_{(m)}\}$;
Search $k$ according to Eq. (6);
if $k$ does not exists, then
    output: $\mathbf{x}^*$ is an ID sample;
else
    output: the $k$ pre-trained models corresponding to the $p$-values $p_{(1)} \cdots p_{(k)}$;
    output: $\mathbf{x}^*$ is an OOD sample.
end if
```

3.3. Sample-Aware Model Selection for OOD Detection

We utilize the Benjamini-Hochberg procedure [2] and devise a threshold adjustment scheme for sample-aware model selection. The ultimate OOD detection decision is derived from the model selection outcome. The Benjamini-Hochberg procedure operates on the following principles.

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 $\mathbf{x}^*$, we compute the score value $\mathbf{s}_j^* = S(\mathbf{x}^*; \phi_j)$ and the p-value p_j for $\phi_j \in \mathcal{M}$, and then sort the p-values in ascending order: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. We then adjust the threshold for the p-value $p_{(j)}$ as $\frac{j}{m}\alpha$ rather than α. In Section 4, we demonstrate that this adjustment maintains TPR at level $1 - \alpha$. Next, we identify the largest subscript that satisfies the threshold condition:

$$k = \max\big\{j : p_{(j)} \leq \frac{j}{m}\alpha\big\}. \tag{6}$$

The k pre-trained models corresponding to the p-values $p_{(1)} \cdots p_{(k)}$ are selected as the active models that can detect the test input as an OOD sample. If k does not exist, i.e. $\big\{j : p_{(j)} \leq \frac{j}{m}\alpha\big\} = \emptyset$, then the active set $\mathcal{A}(\mathbf{x}^*; \mathcal{M}) = \emptyset$ and the test input is classified as an ID sample. We refer to this method as Zoo-based OOD Detection Enhancement (ZODE) and provide the details of ZODE in Algorithm 1.

4. Theoretical Analysis

In this section, we provide a theoretical analysis of the true positive rate (TPR) and false positive rate (FPR) associated with Algorithm 1. Our objective is to establish formal guarantees that demonstrate the reliability and effectiveness of our method. To accomplish this, we develop a rigorous theoretical framework that allows us to derive a bound on the TPR and prove the convergence of the FPR as the number of pre-trained models in the model zoo increases.

Theorem 1 *Suppose that we have access to a pre-trained model zoo denoted by $\mathcal{M} = \{\phi_1, \phi_2, \ldots, \phi_m\}$ and let the target TPR level be $1-\alpha$ with $\alpha \leq 0.5$. If the test input $\mathbf{x}^*$ is an ID sample that $\mathbf{x}^* \sim P_{\mathbf{x}}$ and $\mathbf{s}_j^* = S(\mathbf{x}^*; \phi_j)$ is independent of $\mathbf{s}_{j'}^* = S(\mathbf{x}^*; \phi_{j'})$ for $\forall j \neq j'$, then Algorithm 1 can identify $\mathbf{x}^*$ as an ID sample with probability not less than $1-\alpha$.*

In Theorem 1, we assume that the scoring output of pre-trained model ϕ_j on input $\mathbf{x}^*$, denoted as $\mathbf{s}_j^* = S(\mathbf{x}^*; \phi_j)$, is independent of the scoring output of pre-trained model $\phi_{j'}$ on the same input, denoted as $s_{j'}^* = S(\mathbf{x}_*; \phi_{j'})$. This assumption implies the independence between p_j and $p_{j'}$ for all $j \neq j'$. This assumption holds true when the pre-trained models in the model zoo learn distinct features. In such cases, the model zoo exhibits the desired diversity of features, enabling Algorithm 1 to accurately identify in-distribution (ID) samples.

However, in practice, the pre-trained models may exhibit significant diversity, yet different models might extract related features. In such scenarios, the assumption of independence may not hold, and the p-values associated with different models may be correlated. To overcome this challenge, we present the empirical true positive rate (TPR) of our method ZODE in Section 5. Our experimental results illustrate that ZODE can consistently maintain an empirical TPR no less than the target level, even when the p-values are correlated.

Next, we conduct an asymptotic analysis of the false positive rate (FPR) of Algorithm 1 as the number of pre-trained models in the model zoo, denoted as m, approaches infinity. By examining the FPR in this manner, we can develop a more comprehensive understanding of the correlation between the number of models in the zoo and the algorithm's efficacy in detecting OOD samples.

Theorem 2 *Assuming an OOD sample $\mathbf{x}^* \sim Q$, we consider a fixed proportion π of pre-trained models capable of recognizing $\mathbf{x}^*$ as an OOD sample. We further assume, for any $0 \leq u \leq 1$,*

$$G(u) = \mathbb{P}\big(p_j \leq u | \phi_j \in \mathcal{A}(\mathbf{x}^*; \mathcal{M})\big),$$

where $\mathcal{A}(\mathbf{x}^*; \mathcal{M})$ refers to the set of active models that classify $\mathbf{x}^*$ as OOD, i.e.,

$$\mathcal{A}(\mathbf{x}^*; \mathcal{M}) = \big\{\phi : \phi \in \mathcal{M}, G(\mathbf{x}^*; \phi) = OOD\big\},$$

and $G(u)$ is a distribution different from the uniform distribution $U[0,1]$ and satisfies $(1-\pi) + \pi G^{'}(0) > \frac{1}{\alpha}$. Then, as the number of pre-trained models approaches infinity, ZODE demonstrates the capability to identify OOD samples with a high probability.

Theorem 2 indicates that when the number of pre-trained models is sufficiently large, Algorithm 1 can effectively detect OOD samples with a high probability.

5. Experiments

This section provides an empirical evaluation of the efficacy of our proposed method. We conduct experiments to investigate whether our model zoo and sample-aware model selection scheme can enhance the performance of OOD detectors. Additionally, we demonstrate that ZODE effectively utilizes the diversity of pre-trained models and harnesses the complementarity among single-model detectors, resulting in superior performance. Finally, we present evidence suggesting that ZODE can significantly improve upon the baseline results.

Dataset: We evaluate our proposed method using the CIFAR benchmarks. Specifically, we utilize CIFAR10 [15] as the ID dataset and assess the performance of out-of-distribution (OOD) detectors on six OOD datasets: SVHN [22], LSUN [37], iSUN [34], Texture [4], Places365 [38], and CIFAR100 [15]. To further evaluate the effectiveness of our proposed method, we consider more challenging benchmarks based on ImageNet. Here, we employ ImageNet-1K [6] as the ID dataset and evaluate OOD detectors on four test datasets that are subsets of Places365 [38], iNaturalist [31], SUN [33], and Texture [4]. These datasets contain different categories compared to the ID dataset, rendering them more challenging for OOD detection.

Metrics: Our evaluation of the OOD detection methods is based on three metrics: (1) the true positive rate (TPR) of the ID samples, (2) the false positive rate (FPR) of OOD samples when the TPR of the ID samples reaches approximately 95%, and (3) the area under the receiver operating characteristic curve (AUC). FPR and AUC are widely used in the literature to assess the performance of OOD detectors. To compute the AUC metric, we employ a grid of TPR values ranging from 0 to 1, with a step size of $0.5e-3$. Subsequently, we derive the corresponding FPR values and calculate the area under the receiver operating characteristic curve.

Our focus in this study is on OOD detection without any prior exposure to OOD samples. The primary evaluation metric we utilize is the FPR, while the secondary metric is

Table 1. Comparison between the baseline methods and the corresponding ZODE-enhanced detector. The ID dataset is CIFAR10. All values are percentages. ↓ indicates smaller values are better and vice versa.

		OOD Dataset											
Method		**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-Mahalanobis	94.99	18.24	96.30	6.28	98.48	7.17	98.55	3.88	99.12	72.25	85.93	21.56	95.68
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**

the area under the receiver operating characteristic curve. When two detectors exhibit similar TPR, we compare their performance based on the FPR. On the other hand, if two detectors exhibit different TPR levels, we turn to the AUC to evaluate their performance. A higher AUC suggests the presence of a TPR level at which the detector outperforms others in terms of FPR. However, in practical scenarios, it is challenging to determine the optimal TPR level without access to OOD samples.

Enhanced OOD detection: We consider four different kinds of OOD detection scores: MSP [11], Energy [19] (based on logits) , as well as Mahalanobis [16] and KNN [28] (quantifying the distance in the embedding space). We take them as the baseline methods and denote our enhanced methods by 'ZODE-MSP', 'ZODE-Energy', 'ZODE-Mahalanobis', and 'ZODE-KNN' respectively.

5.1. Evaluation on CIFAR10 benchmarks

Model Zoo. We constructed a model zoo consisting of seven pre-trained models, namely ResNet18, ResNet34, ResNet50, ResNet101, ResNet152 [10], DenseNet [13], and ResNet18* [28]. ResNet and DenseNet are two commonly used backbones in the literature on OOD detection. To ensure diversity in our model zoo, we included six models trained using different architectures and the cross-entropy loss. Moreover, we investigated the impact of the loss function and introduced the ResNet18* model trained with contrastive loss. In summary, our model zoo exhibits diversity in terms of the architectures used and the training strategies employed. This diversity is crucial in evaluating the generalization performance of our proposed method across different models and training regimes.

ZODE maintains TPR. As discussed in Section 3.1, controlling the true positive rate of the ID data is a critical challenge in model selection. Theorem 1 provides theoretical guarantees that if different pre-trained models capture distinct features, ZODE can effectively maintain the TPR close to the target level. In Table 1, we present the empirical TPR of ZODE, which closely approximates the target level of 95%. This result confirms the effectiveness of our proposed method in accurately detecting ID samples.

ZODE achieves consistent improvements. Based on the results presented in Table 1, we can observe consistent performance improvements of ZODE-enhanced detectors compared to their respective baselines. To ensure a fair comparison, we set $k = 50$ in the experiments of ZODE-KNN, which is the same as in Sun et al. [28]. Our results show that compared to the best baseline, KNN+, ZODE-KNN reduces the FPR from 11.07% to 3.83%, which represents a significant improvement in the relative detection accuracy of 65.40%. These findings highlight the effectiveness of our proposed method in enhancing the detection performance of existing OOD detection methods.

ZODE leverages the complementarity between the single-model detectors. Table 2 presents the results of all single-model detectors derived from our model zoo and KNN score. It is evident from the results that the ZODE-KNN detector significantly outperforms all single-model KNN detectors. Specifically, compared with the best single-model baseline, ZODE reduces the FPR from 11.03% to 3.83%, which represents a significant improvement in the relative detection accuracy of 65.28%. These results suggest that the superior performance of ZODE cannot be attributed solely to any single-model detector, and therefore, our proposed method is necessary for the observed improvements. Overall, the results highlight the effectiveness of our proposed method in enhancing the detection performance of existing OOD detection methods.

Table 2. Compare ZODE-KNN detector with single-model KNN detectors. The ID dataset is CIFAR10. All values are percentages. ↓ indicates smaller values are better and vice versa.

		OOD Dataset											
Method		**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
DenseNet	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**

Evaluations on CIFAR10 vs CIFAR100. We tackle a challenging OOD detection task that involves identifying OOD samples drawn from CIFAR100 when the ID data is CIFAR10. Table 3a provides a detailed comparison of our proposed method with competitive OOD detection methods, namely GRAM ([25]), MaSF ([9]), SSD ([26]), and KNN ([28]). Our results show that compared with the best baseline, SSD+, ZODE-KNN reduces the FPR by 20.21%, which represents a relative improvement in detection power of 52.49%. These findings demonstrate the effectiveness of our proposed method in accurately detecting OOD samples in a challenging setting. Furthermore, Table 3b highlights the superiority of ZODE over single-model-based KNN detectors. Our ensemble scheme effectively leverages the complementarity between the single-model detectors, leading to significant improvements in overall detection accuracy.

5.2. Evaluation on ImageNet benchmarks

Model zoo and implementation details. To construct a diverse model zoo, we utilize five pre-trained models with varying architectures and pre-training strategies. Our models include ResNet50* [28], semi-weekly supervised ResNeXt101 32x16d [35], Dinov2-VitL14[5, 23] as well as Swinv2-B256, Swinv2-B384, and Swinv2-L256 [20], with the latter three models having resolutions of 256x256, 256x256, and 384x384, respectively. ResNet50* is trained using the SupCon loss [14], which effectively pulls together points belonging to the same class in the embedding space while separating samples from different classes. ResNeXt101 is pre-trained on Billion-scale images with meta information semantically relevant to ImageNet, achieving an impressive 84.8% top-1 accuracy on ImageNet. The three Swinv2 models are pre-trained at higher resolution, with all achieving top-1 accuracy on ImageNet exceeding 84%. DinoV2-VitL14 is a deep learning model based on self-supervised learning that achieves an accuracy of 83.8% on the ImageNet. For our subsequent analysis, we exclusively report results based on the ZODE-KNN method using our model zoo. We set the hyperparameter α to 6.50% , which yields an empirical TPR for ZODE-KNN of around 95%. For ResNet50*, we utilize a value of $k = 1000$, consistent with Sun et al. [28]. However, for the other models, we select the value of k from the set $100, 200, 500, 700, 800, 900, 1000,$ chosen specifically to minimize the FPR.

ZODE+KNN achieves superior performance. Table 4 presents a comparison of ZODE-KNN with a range of competitive baseline OOD detection methods, including MSP [11], ODIN [18], Energy [19], GODIN [12], Mahalanobis [16], KNN [28], SSD+ [26], and KNN+ [28]. ZODE-KNN outperforms the leading baseline method, KNN+, by reducing the average false positive rate (FPR) from 38.47% to 24.14%. This represents a significant improvement in detection power, with a relative enhancement of 37.25%. Notably, ZODE-KNN demonstrates superior performance on challenging test datasets, such as iNaturalist and Textures, reducing the relative FPR by 92.48% and 79.61%, respectively. These results highlight the effectiveness of ZODE in improving the accuracy and robustness of OOD detection.

ZODE combines the advantages of the single-model detectors. Table 5 presents the performance of every single-model detector derived from our model zoo. We observe three notable trends: (1) ZODE-KNN outperforms the best single-model KNN detector by a relative improvement of 10.49% in FPR, highlighting the efficacy of ZODE on the ImageNet benchmarks, and the significance of the sample-aware model selection scheme. (2) ZODE combines the advantages of different single-model detectors, as evidenced by the varied performance of ResNet50* and ResNeXt101 32x16 on Textures and iNaturalist, and the complementary performance of Swin models on these datasets. This demonstrates that the ZODE-enhanced detector achieves strong and stable performance in all test datasets. (3) ZODE leverages the complementarity between the single-model detectors.

Table 3. CIFAR10 vs CIFAR100. The ID dataset is CIFAR10. All values are percentages. ↓ indicates smaller values are better and vice versa.

(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) ZODE vs Single-model. The detection score is KNN distance.

Model	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**

Table 4. Comparison with baseline methods. The ID data is ImageNet-1K. All values are percentages. ↓ indicates smaller values are better and vice versa.

		OOD Dataset									
Method		**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	94.71	**2.27**	**99.09**	**41.74**	**91.29**	**49.37**	**88.88**	**3.17**	**99.12**	**24.14**	**94.59**

To further illustrate the effectiveness of ZODE, we take Textures as an example to demonstrate how ZODE exploits the diversity of multiple pre-trained models. At Step 6 of Algorithm 1, if $p_{(1)} \leq \frac{1}{m}\alpha$ and $p_{(j)} > \frac{j}{m}\alpha, \forall j \geq 2$, i.e. $k = 1$, then there is only one pre-trained model that can help to identify the test input as an OOD sample. Figure 1 showcases five images from Texture, where each image corresponds to a pre-trained model that successfully identifies it as an OOD sample, while the other models fail to classify it accordingly.

5.3. Compare with related algorithm

In this subsection, we conduct a comparative analysis between our proposed method and a multiple-testing framework [21], which employs a combination of various test statistics using the Benjamini-Yekutieli correction [3]. We extend their method to the model zoo setting and perform a comparative evaluation of our proposed method, ZODE, against the multiple-testing approach. Additionally, this paper compares three common ensemble methods for out-of-distribution (OOD) detection: the Naive detection framework (discussed in Section 3.1), the Average detection framework (which calculates the average p-value obtained from all models in the model zoo for OOD detection), and the Voting detection framework (where an input is classified as an OOD sample if it is deemed OOD by at least 60% of the models in the model zoo). The experimental results are presented in Table 6.

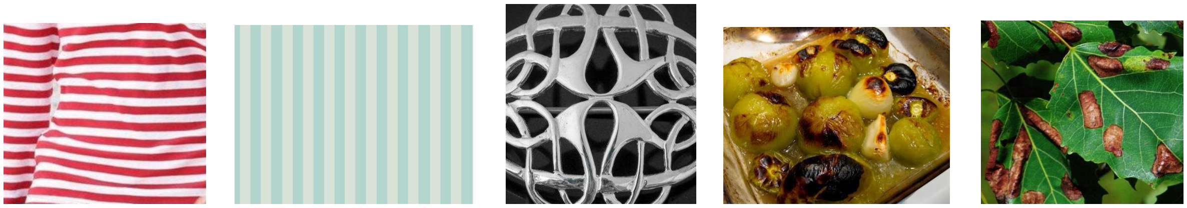
The experimental results demonstrate that ZODE achieves a well-controlled TPR to the target level while maintaining a low FPR. In contrast, the Naive method fails to maintain the target TPR level, and its low FPR is unreliable, as predicted by our theoretical analysis in Section 3.1. The Average and Voting schemes exhibit high TPRs but also high FPRs. On the other hand, the Multiple scheme is more conservative in TPR control and has a higher FPR than ZODE. Overall, these results emphasize the effectiveness of our proposed method in accurately detecting OOD samples while maintaining a low FPR.

6. Conclusion and Limitation

This study aims to enhance the performance of post hoc OOD detection by harnessing the diversity of multiple pre-trained models in a model zoo. To accomplish this objective, we propose ZODE, a novel sample-aware model selection scheme for OOD detection. Extensive exper-

Table 5. Compare ZODE-KNN detector with single-model KNN detectors. The ID dataset is ImageNet-1K. All values are percentages. ↓ indicates smaller values are better and vice versa.

		OOD Dataset									
Method		**iNaturalist**		**SUN**		**Places**		**Textures**		**Average**	
	TPR	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
ResNet50*	95.00	30.52	94.87	48.70	89.03	58.78	85.23	15.46	95.54	38.37	91.17
ResNext101 32x16	95.00	15.11	96.79	55.85	88.62	61.54	86.29	25.99	93.54	39.62	91.31
Swinv2-B256	95.00	9.30	97.91	58.09	88.79	58.45	87.13	41.33	89.68	41.79	90.88
Swinv2-B384	95.00	5.65	98.51	49.66	90.30	52.03	88.50	38.39	89.98	36.43	91.82
Swinv2-L256	95.00	7.03	98.44	51.98	89.54	53.55	88.10	39.15	89.90	37.93	91.49
Dinov2-VitL14	95.00	3.84	99.02	**29.84**	**92.66**	**38.47**	**89.84**	35.74	90.81	26.97	93.08
ZODE-KNN	94.71	**2.27**	**99.09**	41.74	91.29	49.37	88.88	**3.17**	**99.12**	**24.14**	**94.59**



(a) ResNet50* (b) ResNeXt101 (c) Swinv2-B256 (d) Swinv2-B384 (e) Swinv2-L256

Figure 1. We consider an OOD detection task with ImageNet as ID data. This figure presents five OOD images from Texture that only one single-model detector can identify while the other four models fail to classify it as an OOD sample. The detection score is the KNN distance [28].

Table 6. Compare ZODE with four ensemble schemes. The ID dataset is ImageNet-1K. All values are percentages. ↓ indicates smaller values are better and vice versa.

		OOD Dataset									
Method		**iNaturalist**		**SUN**		**Places**		**Textures**		**Average**	
	TPR	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑	FPR↓	AUC↑
Naive	85.76	0.62	99.06	16.59	91.20	23.99	88.50	0.46	99.10	10.42	94.46
Average	98.06	9.86	98.99	56.34	**92.42**	60.79	**90.07**	50.67	94.17	44.42	93.91
Voting	95.41	4.35	98.98	45.51	91.35	50.23	89.09	19.26	96.48	29.84	93.98
Multiple	98.23	8.76	99.05	72.94	90.94	74.66	87.84	12.48	99.10	42.21	94.24
ZODE	**94.71**	2.27	**99.09**	41.74	91.29	49.37	88.88	3.17	**99.12**	24.14	**94.59**

iments demonstrate that ZODE effectively addresses the missed detection problem encountered by single-model detectors by leveraging the complementarity of multiple detectors. Specifically, our findings indicate that combining ZODE with the K-nearest neighbors (KNN) detector yields promising results.

However, it is worth noting that ZODE has a limitation in that it requires a significant amount of storage space during the testing stage. This is because traditional post hoc OOD detection methods only need to pass the threshold from the training stage to the testing stage, while ZODE needs to calculate p-values using the score values of validation samples. Consequently, the testing stage of ZODE necessitates more storage space compared to conventional post hoc OOD detection methods. Nevertheless, this issue can be mitigated through distributed computing.

Acknowledgements

This work was supported in part by the National Nature Science Foundation of China (No.12201048), National Natural Science Foundation of China (No.12071119). The authors thank support from the Interdisciplinary Intelligence Super Computer Center of Beijing Normal University at Zhuhai.

References

[1] Felix Abramovich and Ya'acov Ritov. *Statistical theory: a concise introduction*. CRC Press, 2013. 3

[2] 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. 2, 3

[3] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001. 7

[4] 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*, pages 3606–3613, 2014. 4

[5] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023. 6

[6] 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*, pages 248–255. Ieee, 2009. 4

[7] 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. 1

[8] Mark S Graham, Walter HL Pinaya, Petru-Daniel Tudosiu, Parashkev Nachev, Sebastien Ourselin, and Jorge Cardoso. Denoising diffusion models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2947–2956, 2023. 1

[9] 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. 6

[10] 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*, pages 770–778, 2016. 5

[11] 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. 1, 5, 6

[12] 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*, pages 10951–10960, 2020. 6

[13] 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*, pages 4700–4708, 2017. 5

[14] 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. 6

[15] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 4

[16] 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. 1, 5, 6

[17] 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*, pages 11578–11589, 2023. 1

[18] 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. 1, 6

[19] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020. 1, 5, 6

[20] 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. 6

[21] Akshayaa Magesh, Venugopal V Veeravalli, Anirban Roy, and Susmit Jha. Multiple testing framework for out-of-distribution detection. *arXiv preprint arXiv:2206.09522*, 2022. 7

[22] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 1, 4

[23] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 6

[24] Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, and Andrew Beng Jin Teoh. Understanding the feature norm for out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1557–1567, 2023. 1

[25] Chandramouli S Sastry and Sageev Oore. Zero-shot out-of-distribution detection with feature correlations, 2020. 6

[26] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. In *International Conference on Learning Representations*, 2020. 1, 6

[27] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, 2021.

[28] 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*, pages 20827–20840. PMLR, 2022. 1, 5, 6, 8

[29] Jihoon Tack, Sangwoo Mo, Jongheon Jeong, and Jinwoo Shin. Csi: Novelty detection via contrastive learning on dis-

tributionally shifted instances. *Advances in neural information processing systems*, 33:11839–11852, 2020. 1

[30] Aad W Van der Vaart. *Asymptotic statistics*. Cambridge university press, 2000. 3

[31] 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*, pages 8769–8778, 2018. 4

[32] Guoxuan Xia and Christos-Savvas Bouganis. On the usefulness of deep ensemble diversity for out-of-distribution detection. *arXiv preprint arXiv:2207.07517*, 2022. 1

[33] 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. 4

[34] 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. 4

[35] 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. 6

[36] Donghun Yang, Kien Mai Ngoc, Iksoo Shin, Kyong-Ha Lee, and Myunggwon Hwang. Ensemble-based out-of-distribution detection. *Electronics*, 10(5):567, 2021. 1

[37] 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. 4

[38] 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. 4