Enhancing Out-of-Distribution Detection with Multitesting-based Layer-wise Feature Fusion

Jiawei Li[1*], Sitong Li[1*†], Shanshan Wang[1*], Yicheng Zeng[2], Falong Tan[3], Chuanlong Xie[1]
[1]Beijing Normal University, China
[2]Shenzhen Research Institute of Big Data, The Chinese University of Hong Kong (Shenzhen), China
[3]Hunan University, China
{ljw2420703487, shannyla}@163.com, LST050505@126.com,
statzyc@sribd.cn, falongtan@hun.edu.cn, clxie@bnu.edu.cn

***Abstract*—Deploying machine learning in open environments presents the challenge of encountering diverse test inputs that differ significantly from the training data. These out-of-distribution samples may exhibit shifts in local or global features compared to the training distribution. The machine learning (ML) community has responded with a number of methods aimed at distinguishing anomalous inputs from original training data. However, the majority of previous studies have primarily focused on the output layer or penultimate layer of pre-trained deep neural networks. In this paper, we propose a novel framework, Multitesting-based Layer-wise Out-of-Distribution (OOD) Detection (MLOD), to identify distributional shifts in test samples at different levels of features through rigorous multiple testing procedure. Our approach distinguishes itself from existing methods as it does not require modifying the structure or fine-tuning of the pre-trained classifier. Through extensive experiments, we demonstrate that our proposed framework can seamlessly integrate with any existing distance-based inspection method while efficiently utilizing feature extractors of varying depths. Our scheme effectively enhances the performance of out-of-distribution detection when compared to baseline methods. In particular, MLOD-Fisher achieves superior performance in general. When trained using KNN on CIFAR10, MLOD-Fisher significantly lowers the false positive rate (FPR) from 24.09% to 7.47% on average compared to merely utilizing the features of the last layer.**

***Index Terms*—Out-of-Distribution Detection, Multiple Hypothesis Testing, Feature Fusion**

I. Introduction

Many deep learning systems have achieved state-of-the-art recognition performance when the training and testing data are identically distributed. However, neural networks make high-confidence predictions even for inputs that are completely unrecognizable and outside the training distribution [49], leading to a significant decline in prediction performance or even complete failure. Therefore, the detection of out-of-distribution testing samples is of great significance for the safe deployment of deep learning in real-world applications. This detection process determines whether an input is In-Distribution (ID) or Out-of-Distribution (OOD). OOD detection has been widely utilized in various domains, including medical diagnosis [45] , video self-supervised learning [53] and autonomous driving [6].

Recent advancements in representation learning have led to the development of distance-based OOD detection methods. These methods map a testing input into a suitable feature space and utilize a distance-based score function to determine if the testing input belongs to the ID or OOD category based on its relative distance to the training data [35], [54], [58], [60]. These methods commonly depend on a pre-trained encoder, which maps the test input to an embedding space while preserving the dissimilarity between the test input and the training data. Typically, the pre-trained encoder is a sub-network extracted from a pre-trained classifier, with most existing methods employing feature mapping from the input layer to the penultimate layer. These extracted features are generally considered as high-level semantic features that exhibit strong relevance to the corresponding labels.

However, existing methods tend to overlook the feature representations extracted in shallow layers. In this work, we argue that these low-level features, which capture local and background information, might contain valuable and crucial information for reflecting the dissimilarity between the test input and the training data. We formulate the OOD detection task as a hypothesis testing problem:

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

Here $P_{\mathbf{x}}$ is the training distribution, $\mathcal{Q}$ is a set of distributions and $P_{\mathbf{x}}$ is not included in $\mathcal{Q}$. In the open world scenario, the distributions within the set $\mathcal{Q}$ exhibit diversity, and changes in the distribution between $P_{\mathbf{x}}$ and $Q \in \mathcal{Q}$ can occur at any level of features, including both high-level semantic features and low-level localized features. Consequently, fully leveraging the features extracted from different layers of the neural network can provide a wealth of comprehensive signals to aid the out-of-distribution detection method in identifying distributional shifts.

Several studies have highlighted the effectiveness of utilizing multi-scale features extracted from different intermediate layers for OOD detection [18], [39], [72]. For example, MOOD [39] adaptively selects intermediate classifier outputs for OOD inference based on the complexity of the test inputs. However, in the case of MOOD, the primary motivation for adaptively selecting the optimal exit is to reduce computational costs rather than enhance the OOD detection accuracy. In an-

[*]These authors contributed equally.

[†]Correspondence to: LST050505@126.com

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00102

other recent study [18], the authors propose treating the scores computed on the features of each layer as a type of functional data and identifying out-of-distribution samples by integrating changes in functional trajectories. Nevertheless, a potential issue with this approach is that not all features extracted from the intermediate layers are relevant for OOD detection. There are multiple options available for aggregating the feature scores extracted from each layer. These options include selecting feature similarity scores for each layer and determining the metric for score trajectory differences. However, the selection and determination of these aggregation methods are currently open issues in the field. The Multi-scale OOD detection (MODE, [72]) is an attention-based method that utilizes both global visual information and local region details of images to enhance OOD detection. It introduces a trainable objective called Attention-based Local Propagation, which utilizes a cross-attention mechanism to align and emphasize the local regions of the target objects in pairwise examples. However, the aforementioned methods necessitate modifications in the pre-training method or the backbone network, as well as the selection of similarity metrics for the score trajectories.

In this paper, we propose a novel and general framework, called **M**ultitesting-based **L**ayer-wise **O**ut-of-Distribution **D**etection (**MLOD**) to enhance the performance of detecting samples that the MLOD has not encountered during training. Our proposed approach utilizes commonly used pre-training methods and models and leverages p-values to normalize the detection output at each layer. It aims to determine whether there exists a layer of features in a multi-layered pre-trained MLOD that can effectively detect distributional shifts between the test sample and the training data. To accomplish this, our approach calculates p-values based on the empirical distribution of the score function across different layers and employs multiple hypothesis testing techniques to control the True Positive Rate (TPR). Additionally, our framework can identify the layer that can detect the presence of distributional shifts between the test sample and the training data. Considering the potential high correlation between features extracted from different layers of a pre-trained neural network, we adopt five multiple hypothesis testing methods to adjust the p-value. These methods include the Benjamini-Hochberg procedure [8], adaptive Benjamini-Hochberg procedure [9], Benjamini-Yekutieli procedure [10], Fisher's method [16], and Cauchy combination test [41]. We conduct systematic experimental comparisons to illustrate the practical advantages of MLOD on several benchmarks. On CIFAR10, the MLOD-Fisher method significantly reduces the False Positive Rate (FPR) from 24.09% to 7.47% on average and consistently outperforms the other methods on five OOD datasets.

Our main contributions are summarized as follows:

- We propose a novel OOD detection framework from the perspective of the multi-layer feature of deep neural networks, namely **M**ultitesting-based **L**ayer-wise **O**ut-of-distribution **D**etection.
- We provide a comprehensive evaluation of the effectiveness of **MLOD** through both theoretical understanding and experimental verification, focusing on multiple combinatorial tests. The main multiple test methods that we consider in our evaluation include BH, adaptiveBH, BY, Fisher method, and Cauchy method.
- Extensive experiments demonstrate that MLOD outperforms post-hoc methods that solely rely on the feature of the final output layer, as well as enhance the performance of various existing OOD scores. These experiments were conducted using the current benchmarks, and the results indicate a significant improvement in performance.

II. Preliminaries

The primary aim of OOD detection is to ascertain whether a given input is sampled from the training distribution or not. Let $\mathcal{X}$ and $\mathcal{Y}$ denote the input and label space, respectively. The training distribution over $\mathcal{X} \times \mathcal{Y}$ is denoted as $\mathcal{P}_{id}$, while the marginal distribution on $\mathcal{X}$ is denoted as $\mathcal{D}_{id}$. After training a neural network on the training data derived from $\mathcal{P}_{id}$, the feature extractor is represented as $\phi(\mathbf{x})$, which is a sub-network of the pre-trained neural network. The feature-based OOD detector uses a decision function to determine whether a test input belongs to the ID or OOD sample. The decision function is defined as follows:

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

Here, $\mathbf{x}^*$ denotes the test input, and $S(\mathbf{x}^*, \phi)$ is a scoring function to quantify the similarity between the test input $\mathbf{x}^*$ and the training data on the embedding space derived by the feature extractor ϕ. The threshold λ_ϕ acts as a tuning parameter that regulates the probability of misclassifying an ID sample, also known as the True Positive Rate (TPR). To maintain TPR a desired level of $1 - \alpha$, the threshold λ_ϕ is selected based on the α-quantile of the empirical distribution of $\{S(\mathbf{x}_i, \phi)\}_{i=1}^n$. This is given by

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

where $\mathbb{I}\{\cdot\}$ represents the indicator function, and $\{\mathbf{x}_i\}_{i=1}^n$ corresponds to a validation set consisting of n ID inputs. Therefore, $\lambda_\phi = \hat{F}^{-1}(\alpha; \phi) = \inf_{s \in \mathbb{R}} \{s : \hat{F}(s; \phi) \geq \alpha\}$.

III. Challenges

Consider a pre-trained neural network f with m layers. The network can be represented as the composition of functions:

$$f(\mathbf{x}) = h \circ g_m \circ \cdots \circ g_2 \circ g_1(\mathbf{x}), \tag{4}$$

In this equation, h represents the top classifier which operates on the features extracted from the m-th layer. Similarly, g_1 denotes the feature mapping function from the input to the first layer of features. For $2 \leq i \leq m$, g_i is the transformation function from the $(i-1)$-th layer to the i-th layer. We introduce the notation $\phi_1(\mathbf{x}) = g_1(\mathbf{x})$, and define the mapping function from the input layer to the output of the i-th layer as:

$$\phi_i(\mathbf{x}) = g_i \circ \cdots \circ g_1(\mathbf{x}). \tag{5}$$

In the case of OOD samples, distribution shifts can arise at any feature layer. Consequently, we can leverage the hierarchical structure of the MLOD f across its layers to maximize its potential for OOD detection. This allows us to reformulate OOD detection as a layer-wise assessment of similarity between the test input and the training data.

A naive approach to implement layer-wise detection is expressed as follows:

$$G(\mathbf{x}^*; f) = \begin{cases} \text{ID} & \text{if } S(\mathbf{x}^*, \phi_i) \geq \lambda_{\phi_i}, \forall \phi_i; \\ \text{OOD} & \text{if } S(\mathbf{x}^*, \phi_i) < \lambda_{\phi_i}, \exists \phi_i. \end{cases} \tag{6}$$

In this approach, the test input $\mathbf{x}^*$ is classified as ID only if all detectors $G(\mathbf{x}^*, \phi_i)$ agree that $\mathbf{x}^*$ is an ID sample. Conversely, if there exists a layer of features $\phi_i(\mathbf{x}^*)$ for which $G(\mathbf{x}^*, \phi_i) =$ OOD, then $\mathbf{x}^*$ is determined to be an OOD sample.

However, this simple approach suffers from a significant drawback: the True Positive Rate (TPR) deteriorates, rendering the final outcome of the layer-wise OOD detection unreliable. It is important to note that each detector $G(\mathbf{x}^*, \phi_i)$ has a probability α of misclassifying an ID sample as an OOD sample. When aggregating the results from multiple layers, the probability of committing this error accumulates. Specifically, the probability can be expressed as $1 - (1-\alpha)^m$, assuming that the feature mappings $\{\phi_i\}_{i=1}^m$ are independent. As the number of layers in the pre-trained neural network increases, the TPR can approach zero. This observation indicates that the detector presented in Equation (6) fails to maintain the TPR at the desired level.

IV. Methodology

To address the aforementioned issues and challenges, our proposed detection framework aims to achieve the following objectives:

1. Applicability to general pre-trained models: The framework should be applicable to a wide range of pre-trained models without the need for re-training or the use of specialized model architectures.
2. Standardization of layer scores: The framework should ensure that the scores from each layer are standardized, avoiding any bias in decision-making caused by the varying ranges of score distributions in the intermediate layers.
3. Fusion of layer results while maintaining desired TPR level: The framework should be able to effectively fuse the results from each layer while ensuring that the TPR remains within the desired range.

To achieve Objective 1, our framework is designed to be compatible with various types of detection scores, including output-based scores, logits-based scores, and feature-based scores. By accommodating these different types of detection scores, our framework ensures its applicability to a wide range of pre-trained models without the need for re-training or the use of specific model architectures.

To achieve Objective 2, we employ the p-value [2] to standardize the distribution of scores across different layers. For a test input $\mathbf{x}^*$, its p-value is defined by

$$p = P(S(\mathbf{x}, \phi) \leq S(\mathbf{x}^*, \phi) | \mathbf{x} \sim \mathcal{D}_{id}) \tag{7}$$

In fact, when the sample size n is large enough, the decision rule $\{\mathbf{x} : S(\mathbf{x}, \phi) < \lambda_\phi\}$ in Equation (2) is equivalent to the decision rule $\{\mathbf{x} : \text{p-value of } \mathbf{x} < \alpha\}$.

Proposition 1. *For a given input $\mathbf{x}^*$, using the p-value is equivalent to using the hard threshold $S(\mathbf{x}^*) < \lambda$.*

Sketch of Proof: We denote $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ as validation data drawn from the ID distribution, and sort their detection scores in an ascending order: $S_{(1)} \leq S_{(2)} \leq ... \leq S_{(n)}$. Since the threshold λ_ϕ is chosen to guarantee $1 - \alpha 1$ TPR, we have $S_{([\alpha n])} \leq \lambda \leq S_{([\alpha n]+1)}$, where $[\cdot]$ is the floor function. On the other hand, the p-value of $\mathbf{x}^*$ less than 0.05 implies that $P\big(S(\mathbf{x}, \phi) \leq S(\mathbf{x}^*, \phi) \big| \mathbf{x} \sim \hat{\mathcal{D}}_{id}\big) \approx 0.05$ where $\hat{\mathcal{D}}_{id}$ is the empirical distribution of $\{\mathbf{x}_i\}_{i=1}^n$. Hence $S(\mathbf{x}^*, \phi) \lesssim S_{[\alpha n]+1}$.

Proposition 2. *If $\mathbf{x}^*$ is drawn from the ID distribution, the p-value of $\mathbf{x}^*$ follows a uniform distribution $U[0,1]$.*

Sketch of Proof: Let p^* represent the p-value of $\mathbf{x}^*$, $s^* = S(\mathbf{x}^*, \phi)$, and denote $F(s; \phi)$ as the cumulative distribution function of $S(\mathbf{x}, \phi)$ with $\mathbf{x} \sim \mathcal{D}_{id}$. We have

$$p^* = P\big(S(\mathbf{x}, \phi) \leq s^* \big| \mathbf{x} \sim \mathcal{D}_{id}\big) = F(s^*, \phi).$$

By the continuity of $S(\mathbf{x}^*)$ and Lemma 21.1 of [1]:

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

To achieve Objective 3, we employ the technique of multiple hypothesis testing in statistics to adjust the p-value in order to make TPR within the desired target range. We consider five specific methods for multiple hypothesis testing:

- The Benjamini-Hochberg procedure [8]: This procedure controls the false discovery rate (FDR) while controlling the proportion of false positives among the rejected hypotheses.
- The adaptive Benjamini-Hochberg procedure [9]: This procedure is an adaptive version of the Benjamini-Hochberg procedure that provides a more powerful control of the FDR when the number of hypotheses tested is large.
- The Benjamini-Yekutieli procedure [10]: This procedure is a modification of the Benjamini-Hochberg procedure that controls the false discovery rate under arbitrary dependency structures.
- The Fisher's method [16]: This method combines the p-values from multiple hypothesis tests using Fisher's combining function to obtain an overall p-value.
- The Cauchy combination test [41]: This method utilizes the Cauchy combination test to combine p-values from multiple hypothesis tests, providing a robust and powerful approach for multiple hypothesis testing.

TABLE I

RESULTS ON CIFAR10. COMPARISON WITH BASELINE METHODS THAT ONLY UTILIZE THE LAST LAYER FEATURES AND MOOD. THE PRE-TRAINED CLASSIFIER IS MSDNET [24]. THE BEST RESULTS ARE IN BOLD. ALL VALUES ARE PRESENTED AS PERCENTAGES. THE DOWNWARD ARROW INDICATES THAT LOWER VALUES ARE PREFERABLE, AND VICE VERSA.

Detection Score	Method	OOD Dataset											
		SVHN		**LSUN**		**iSUN**		**Texture**		**LSUNR**		**Average**	
		FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑
MSP	Layer@last	54.72	91.43	34.38	95.27	52.27	92.30	**59.15**	88.11	50.49	92.51	50.20	91.92
	MOOD	**53.92**	**91.91**	33.89	95.38	53.29	92.22	60.59	88.01	51.79	92.41	50.70	91.99
	MLOD-BH	61.01	90.77	29.99	**96.04**	47.98	93.50	61.38	89.03	47.43	93.55	49.56	92.58
	MLOD-adaBH	60.82	83.15	31.76	90.02	47.27	89.56	61.49	85.39	46.65	88.01	49.60	87.23
	MLOD-BY	61.03	63.57	**31.09**	67.84	47.88	68.54	61.42	72.33	47.36	66.37	49.76	67.73
	MLOD-Fisher	58.52	91.28	31.91	95.69	46.88	**93.69**	60.72	**89.67**	**45.98**	**93.74**	**48.80**	**92.81**
	MLOD-Cauchy	60.10	91.21	32.70	95.65	**46.59**	93.68	60.51	89.43	46.28	93.71	49.24	92.74
Energy	Layer@last	34.09	92.82	5.91	98.73	33.30	93.59	55.14	82.35	31.57	93.77	32.00	92.25
	MOOD	**30.88**	**93.99**	5.72	98.75	33.78	93.84	58.12	82.48	32.30	93.97	32.16	92.61
	MLOD-BH	45.63	93.22	4.28	99.01	25.85	95.80	58.87	85.79	23.57	96.00	31.64	93.96
	MLOD-adaBH	44.46	86.03	4.19	95.43	23.25	91.43	58.76	79.74	21.93	90.38	30.52	88.60
	MLOD-BY	45.70	62.57	25.29	73.47	29.63	69.12	58.87	64.38	32.01	66.65	38.30	67.24
	MLOD-Fisher	40.02	93.75	3.95	**99.04**	**22.69**	**96.30**	57.92	**86.76**	**21.04**	**96.47**	**29.12**	**94.46**
	MLOD-Cauchy	44.78	93.55	**3.85**	99.03	22.94	96.28	58.26	85.95	21.11	96.44	30.19	94.25
ODIN	Layer@last	39.68	91.10	5.72	98.75	28.62	94.40	**54.36**	81.76	26.79	94.54	31.03	92.11
	MOOD	**35.66**	92.69	5.32	98.79	28.59	94.66	57.23	82.01	27.00	94.76	30.76	92.58
	MLOD-BH	49.95	92.73	**3.46**	99.10	20.21	96.60	56.68	86.22	17.82	96.76	29.62	94.28
	MLOD-adaBH	48.64	85.18	3.74	95.69	18.36	93.02	56.00	80.21	16.30	91.75	28.61	89.17
	MLOD-BY	45.70	62.57	25.29	73.47	29.63	69.12	58.87	64.38	32.01	66.65	38.30	67.24
	MLOD-Fisher	44.59	**93.28**	3.56	**99.11**	**17.38**	**97.05**	55.75	**87.10**	**16.09**	**97.18**	**27.47**	**94.74**
	MLOD-Cauchy	48.86	93.08	3.48	99.10	17.60	97.03	56.21	86.31	16.09	97.15	28.45	94.53

We combine the proposed framework, MLOD, with these five methods, which are denoted as MLOD-BH, MLOD-adaBH, MLOD-BY, MLOD-Fisher, and MLOD-Cauthy respectively.

Recall the difinition of ϕ_i in Equation (5). Given a test input $\mathbf{x}^*$, we compute the score value $S(\mathbf{x}^*; \phi_i)$ and obtain corresponding p-values denote as p_i. After going through all ϕ_i, we obtain m p-values: $\{p_1, ..., p_m\}$. Let the desired level of TPR is $1 - \alpha$. The details of the five methods are described below:

MLOD-BH. We use the idea of the Benjamini-Hochberg procedure. Sort m obtained p-values in ascending order: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. We identify the test input $\mathbf{x}^*$ as an OOD sample if there exists an integer $1 \leq k \leq m$ such that $p_{(k)} \leq \frac{\alpha k}{m}$, otherwise $\mathbf{x}^*$ is classified as an ID sample.

MLOD-adaBH. We sort the p-values in ascending order: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$. If $p_{(i)} \geq \frac{\alpha i}{m}$, $1 \leq i \leq m$, then $\mathbf{x}^*$ is classified as an ID data, otherwise continue to calculate

$$S_i = (1 - p_{(i)})/(m + 1 - i).$$

Set $i = 2$, proceed as $S_i \geq S_{i-1}$ until for the first time $S_j < S_{j-1}$. Then compute $\hat{m_0} = \min([1/S_j + 1], m)$. We identify the test input $\mathbf{x}^*$ as an OOD sample if there exists an integer $1 \leq k \leq m$ such that $p_{(k)} \leq \frac{\alpha k}{\hat{m_0}}$.

MLOD-BY. Based on the idea of Benjamini-Yekutieli procedure, we also sort m obtained p-values in ascending order: $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ and define $k = \max\{i | p_{(i)} \leq \frac{\alpha i}{m f(m)}\}$, where $f(m) = \sum_{i=1}^{m} \frac{1}{i}$. We identify the test input $\mathbf{x}^*$ as an OOD sample if there exists an integer $1 \leq k \leq m$ such that $p_{(k)} \leq \frac{\alpha k}{m}$, otherwise $\mathbf{x}^*$ is classified as an ID sample.

MLOD-Fisher. Fisher's method combines m p-values and constructs a new test statistic: $F = \sum_{i=1}^{m} -2ln(P_i)$. If $F > \chi^2(1-\alpha, 2m)$, then the test input $\mathbf{x}^*$ is classified to be OOD, where $\chi^2(1-\alpha, 2m)$ is the upper α quantile of the chi-square distribution with degrees of freedom of $2m$.

MLOD-Cauchy. Cauchy combination test also combines m p-values and establishes the Cauchy combination test statistic: $T = \sum_{i=1}^{m} w_i \tan{(0.5 - p_i)\pi}$, where the weights w_i's are nonnegative and $\sum_{i=1}^{m} w_i = 1$. If $T > t_{1-\alpha}$, then the test input $\mathbf{x}^*$ is defined to be an OOD sample, where $t_{1-\alpha}$ is the upper α quantile of the standard Cauchy distribution.

V. EXPERIMENTAL SETTING

Datasets. For the evaluation on CIFAR Benchmarks, we utilize CIFAR-10 and CIFAR-100 as the in-distribution datasets, respectively. Additionally, we assess the OOD detector on a total of 5 OOD datasets: LSUN (crop) [71], SVHN [48], Textures [13], iSUN [67], and LSUN (resize) [71]. All images are resized to 32×32.

Models. We ran experiments with three models. A MSDNet [24] pre-trained on ILSVRC-2017 with over 5M parameters and achieves a top-1 accuracy of 75%. A ResNet-18 [19] model with top-1 test set accuracy of 70% and over 11M parameters. A ResNet-34 [19] model with top-1 test set accuracy of 74% and over 21M parameters. We download all the checkpoints weights from PyTorch [50] hub. All models are trained from scratch on CIFAR10 or CIFAR100.

Evaluation Metrics. Our evaluation of OOD detection methods utilizes two metrics: (1) the false positive rate (FPR)

TABLE II

Results on CIFAR. Comparison with baseline methods that only utilize the last layer features. The pre-trained classifier is ResNet-18 [19]. The best results are in bold. All values are presented as percentages. The downward arrow indicates that lower values are preferable, and vice versa.

CIFAR10	**Method**	OOD Dataset											
		SVHN		**LSUN**		**iSUN**		**Texture**		**LSUNR**		**Average**	
		FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑
KNN	Layer@last	27.95	95.49	18.48	96.84	24.65	95.52	26.7	94.97	22.67	96.07	24.09	95.78
	MLOD-BH	10.07	93.23	5.11	85.91	9.69	87.42	20.71	65.27	8.03	90.42	10.72	84.45
	MLOD-adaBH	9.56	93.33	4.70	85.98	9.25	87.48	20.35	65.05	7.62	90.50	10.30	84.47
	MLOD-BY	10.02	93.28	5.02	85.97	9.55	87.40	20.61	64.58	7.99	90.43	10.64	84.33
	MLOD-Fisher	**6.20**	97.32	**1.48**	97.77	**6.61**	97.18	**18.19**	94.66	**4.86**	97.44	**7.47**	96.87
	MLOD-Cauchy	8.52	**98.16**	3.76	**98.72**	8.64	**98.05**	19.8	**95.50**	6.95	**98.35**	9.53	**97.76**

CIFAR100	**Method**	OOD Dataset											
		SVHN		**LSUN**		**iSUN**		**Texture**		**LSUNR**		**Average**	
		FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑	FPR95↓	AUC↑
KNN	Layer@last	56.35	86.49	77.66	78.29	71.11	83.45	67.27	83.31	66.79	85.29	67.84	83.37
	MLOD-BH	47.12	91.20	58.08	91.25	42.32	91.52	34.38	77.94	46.17	91.98	45.61	88.78
	MLOD-adaBH	43.59	91.55	56.42	91.11	38.80	92.17	32.80	77.59	42.33	92.76	42.79	89.04
	MLOD-BY	46.97	90.09	57.96	91.05	42.23	90.84	34.34	76.14	45.91	91.32	45.48	87.89
	MLOD-Fisher	**40.01**	**92.05**	50.34	92.99	25.99	94.73	**30.40**	91.85	**27.07**	**94.74**	**34.76**	**93.27**
	MLOD-Cauchy	43.79	91.96	55.18	92.11	36.63	93.88	32.59	**92.45**	39.94	93.65	41.63	92.81

of OOD data when the true positive rate (TPR) of the in-distribution (ID) data is approximately 95% (referred to as **FPR95**); and (2) the area under the receiver operating characteristic curve (**AUC**).

Detection Score. We consider five OOD detection scores: MSP [20], ODIN [38], Energy [40] and KNN [58]. MSP [20] regards the maximum softmax probabilities as the detection score. ODIN [38] utilizes temperature scaling and adds small perturbations to distinguish the softmax scores between ID and OOD samples. The energy-based model [33] maps a test input to a scalar that is higher for OOD samples and lower for the training data. Liu et al. [40] propose an energy score that utilizes the logits outputted by a pre-trained classifier. KNN [58] is a distance-based detector that utilizes the feature distance between a test input and the k-th nearest ID data. In this paper, we set the hyperparameter k to be 50.

Baselines. In this paper, we consider two baseline approaches. The first approach involves OOD detection using the outputs (probabilities or logits) of pre-trained classifiers or the penultimate layer of features. This approach is commonly used in existing OOD detection methods. By comparing our proposed framework with the approach that solely relies on the information from the last layer, we demonstrate that our method effectively utilizes information from intermediate feature layers to enhance OOD detection performance. The second baseline approach is MOOD [39]. MOOD utilizes a pre-trained model with multiple exits, such as MSDNet, and performs supervised learning of the classification task on all features in the middle layer. In order to improve the sensitivity of OOD detection, we compare our method with MOOD. The comparison reveals that our feature layer selection, which is based on the detection task, outperforms the approach based on input complexity.

VI. Results and Discussion

Main Results. The performance of **MLOD** on CIFAR10 and CIFAR100 benchmarks using ResNet-18 is evaluated. We compare our method against a baseline method that utilizes only the penultimate layer of features. The OOD detection performance for each OOD test dataset, as well as the average across all five datasets, is presented in Table II. Based on the results, **MLOD**-Fisher and **MLOD**-Cauthy consistently outperform the baseline method. We also consider MOOD [39] as a baseline method and employ MSDNet [24] as the pre-trained classifier. In Table I, we compare our method against MOOD and present the OOD detection performance for each OOD test datasets, as well as the average across all five datasets. On the average results, our method outperforms MOOD.

MLOD-Fisher achieves consistent improvements on FPR95. In our comparison of **MLOD** with the baseline method using different multiple testing methods, we observed that**MLOD**-Fisher based on KNN performed better than the baseline methods on average for FPR95. Specifically, we found that the MLOD-Fisher based on KNN method significantly reduced FPR95 from 24.09% to 7.47% on average. This method exhibited consistently better performance for the SVHN, LSUN, iSUN, Texture, and LSUNR datasets.

MLOD-Cauchy achieves consistent improvements on AUC. In our comparison of **MLOD** with baseline methods, we found that**MLOD**-Cauchy based on KNN achieved best performance in terms of AUC. On the other hand, **MLOD**-adaBH and **MLOD**-BY methods have weaker performance in terms of AUC compared to the baseline methods on average. They still offer advantages in terms of their ability to control the false discovery rate and handle arbitrary dependency structures.

MLOD method demonstrates excellent scalability. The

MLOD method is equally applicable in pre-trained deep neural networks with multiple exits. We use default MSDNet [24] with $k = 5$ blocks with 4 layers each and use the default growth rate of 6, with scale factors [1, 2, 4] to apply our approach. The results of CIFAR10 are shown in Table I. On average, compared to the state-of-the-art MOOD method, MLOD-FIsher reduces FPR by 2.8% and improves AUC by 1.39% for CIFAR10 dataset.

MLOD leverages the fusion of multi-layer features. The majority of previous studies have primarily focused on the output layer or penultimate layer of pre-trained deep neural networks. We cite the results of the competitors reported in MOOD [39] and show the average results between MLOD vs. single layer feature and MOOD based on MSDNet with Energy score on 8 OOD datasets in Table III. MLOD uses the feature information of multiple layers of pre-trained deep neural networks to improve performance compared with the single layer.

TABLE III
PERFORMANCE COMPARISON BETWEEN MLOD VS. SINGLE LAYER FEATURE AND MOOD BASED ON MSDNET WITH ENERGY SCORE. NUMBERS ARE AVERAGED RESULTS WITH CIFAR-100 BENCHEMARKS.

	AUC↑	**FPR95**↓
Exit@1	77.69	70.83
Exit@2	83.13	57.19
Exit@3	84.71	57.76
Exit@4	85.31	57.12
Exit@5	84.51	59.15
MOOD	86.21	55.26
MLOD-BH	86.74	52.72
MLOD-adaBH	78.45	52.24
MLOD-BY	57.56	58.40
MLOD-Fisher	**87.19**	**50.84**
MLOD-Cauchy	86.64	51.95

VII. RELATED WORK

The investigation into OOD detection within deep neural networks has been undertaken through diverse perspectives, primarily spanning density-based, distance-based, and classification-based methods. Density-based methods explicitly model in-distribution samples using probabilistic models, flagging test data in low-density regions as OOD. As exemplified in the study by [35], class-conditional Gaussian distributions are utilized to model the distributions of multiple classes within in-distribution samples. [75] introduces a more expressive density function based on deep normalizing flow. Recent approaches have delved into novel OOD scores, with likelihood regret [66] proposing a score applicable to variational auto-encoder (VAE) generative models.

Distance-based methods operate on the premise that OOD samples should exhibit relatively greater distances from the centers of in-distribution samples. In [35], OOD detection is achieved by evaluating the Mahalanobis distance between test samples and their nearest class-conditional distributions. Another non-parametric approach [58] involves computing the k-nearest neighbors (KNN) distances between the embeddings of test inputs and training set embeddings. In addition, several studies make use of the spatial separation between the embedding of the input and the centroids of respective classes [17], [25], [61]. Other works such as SSD+ [52] have adopted off-the-shelf contrastive losses for OOD detection. However, this approach results in embeddings that exhibit insufficient inter-class dispersion. CIDER [42] specifically addresses this issue by optimizing for substantial inter-class margins, thereby yielding more favorable hyperspherical embeddings.

In the domain of classification-based methods, a foundational benchmark was set using maximum softmax probability (MSP) from pre-trained networks [20]. ODIN [37] refined this baseline by integrating temperature scaling and input perturbation to bolster the distinction between ID and OOD samples. Generalized ODIN [22] furthered this progression by introducing a specialized network for learning temperature scaling and a strategy for selecting perturbation magnitudes. Addressing challenges of overconfident posterior distributions in OOD detection using softmax scores, [40] introduced a novel approach utilizing the energy score derived from logit outputs for OOD detection. ReAct [56] proposed a simple and effective technique aimed at mitigating model overconfidence on OOD data. Outlier Exposure (OE) methods [21] utilized a set of collected OOD samples during training to assist the learning of ID/OOD discrepancy. However, such outlier exposure approaches hinge on the availability of OOD training data. In the absence of such data, efforts have been made to synthetically generate OOD samples. VOS [15] proposed the synthesis of virtual outliers from the low-likelihood region in the feature space, which is more tractable given lower dimensionality. Noteworthy contributions from MOS [27] have advocated for OOD detection in large-scale settings, aligning more closely with real-world applications.

Recent research has advanced OOD detection by generating auxiliary datasets or optimizing network structures. HOOD [29] employs an intervention process to generate both malign and benign OOD datasets, enhancing model robustness. DAL [63] reduces distributional differences by forming a set of distributions within Wasserstein spheres. EVIL [28] improves generalization by identifying subnetworks through the exploration of parameters sensitive to distribution changes, demonstrating novel approaches to model resilience and detection capabilities.

VIII. CONCLUSION

In this work, we introduces the MLOD framework, which leverages multitesting-based layer-wise feature fusion for OOD detection. The proposed framework is applicable to various pre-trained models and is supported by a comprehensive experimental analysis that evaluates the performance of various methods. This analysis demonstrates the efficacy of MLOD in improving OOD detection.

REFERENCES

[1] *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

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

[3] Alvin Alpher. Frobnication. *Journal of Foo*, 12(1):234–778, 2002.

[4] Alvin Alpher and Ferris P. N. Fotheringham-Smythe. Frobnication revisited. *Journal of Foo*, 13(1):234–778, 2003.

[5] Alvin Alpher, Ferris P. N. Fotheringham-Smythe, and Gavin Gamow. Can a machine frobnicate? *Journal of Foo*, 14(1):234–778, 2004.

[6] Alexander Amini, Ava Soleimany, Sertac Karaman, and Daniela Rus. Spatial uncertainty sampling for end-to-end control. *arXiv preprint arXiv:1805.04829*, 2018.

[7] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[8] 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.

[9] Yoav Benjamini and Yosef Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of educational and Behavioral Statistics*, 25(1):60–83, 2000.

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

[11] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[12] Chun Fu Chen, Quanfu Fan, Neil Mallinar, Tom Sercu, and Rogerio Feris. Big-little net: An efficient multi-scale feature representation for visual and speech recognition. In *International Conference on Learning Representations*, 2019.

[13] 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.

[14] Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical japanese literature. *arXiv preprint arXiv:1812.01718*, 2018.

[15] Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *arXiv preprint arXiv:2202.01197*, 2022.

[16] Ronald Aylmer Fisher. *Statistical methods for research workers*. Springer, 1992.

[17] Eduardo Dadalto Câmara Gomes, Florence Alberge, Pierre Duhamel, and Pablo Piantanida. Igeood: An information geometry approach to out-of-distribution detection. In *NeurIPS DistShift Workshop 2021*, 2021.

[18] Eduardo Dadalto Câmara Gomes, Pierre Colombo, Guillaume Staerman, Nathan Noiry, and Pablo Piantanida. A functional perspective on multi-layer out-of-distribution detection.

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

[20] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[21] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[22] 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.

[23] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Weinberger. Multi-scale dense networks for resource efficient image classification. In *International Conference on Learning Representations*, 2018.

[24] Gao Huang, Danlu Chen, Tianhong Li, Felix Wu, Laurens van der Maaten, and Kilian Q. Weinberger. Multi-scale dense convolutional networks for efficient prediction. *ArXiv*, abs/1703.09844, 2017.

[25] Haiwen Huang, Zhihan Li, Lulu Wang, Sishuo Chen, Bin Dong, and Xinyu Zhou. Feature space singularity for out-of-distribution detection. *arXiv preprint arXiv:2011.14654*, 2020.

[26] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34:677–689, 2021.

[27] Rui Huang and Yixuan Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8710–8719, 2021.

[28] Zhuo Huang, Muyang Li, Li Shen, Jun Yu, Chen Gong, Bo Han, and Tongliang Liu. Winning prize comes from losing tickets: Improve invariant learning by exploring variant parameters for out-of-distribution generalization. *arXiv preprint arXiv:2310.16391*, 2023.

[29] Zhuo Huang, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, and Tongliang Liu. Harnessing out-of-distribution examples via augmenting content and style. *arXiv preprint arXiv:2207.03162*, 2022.

[30] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[31] G. Larsson, M. Maire, and G. Shakhnarovich. Fractalnet: Ultra-deep neural networks without residuals. In *International Conference on Learning Representations*, 2017.

[32] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[33] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

[34] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570. PMLR, 2015.

[35] 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.

[36] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *ArXiv*, abs/1807.03888, 2018.

[37] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[38] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[39] 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*, pages 15313–15323, 2021.

[40] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020.

[41] Yaowu Liu and Jun Xie. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 2019.

[42] Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? In *The Eleventh International Conference on Learning Representations*, 2022.

[43] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017.

[44] Peyman Morteza and Yixuan Li. Provable guarantees for understanding out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7831–7840, 2022.

[45] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation. *Medical image analysis*, 59:101557, 2020.

[46] Full Author Name. Frobnication tutorial, 2014. Supplied as additional material `tr.pdf`.

[47] Yuval Netzer, Tao Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. 2011.

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

[49] 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*, pages 427–436, 2015.

[50] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An

imperative style, high-performance deep learning library. In *Neural Information Processing Systems*, 2019.

[51] Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

[52] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations (ICLR)*, 2021.

[53] Pritam Sarkar, Ahmad Beirami, and Ali Etemad. Uncovering the hidden dynamics of video self-supervised learning under distribution shifts. *arXiv preprint arXiv:2306.02014*, 2023.

[54] Vikash Sehwag, Mung Chiang, and Prateek Mittal. Ssd: A unified framework for self-supervised outlier detection. *arXiv preprint arXiv:2103.12051*, 2021.

[55] Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21(2):345–352, 2020.

[56] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34:144–157, 2021.

[57] Yiyou Sun and Yixuan Li. Dice: Leveraging sparsification for out-of-distribution detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pages 691–708. Springer, 2022.

[58] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022.

[59] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[60] 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.

[61] Joost Van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. In *International conference on machine learning*, pages 9690–9700. PMLR, 2020.

[62] 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.

[63] Qizhou Wang, Zhen Fang, Yonggang Zhang, Feng Liu, Yixuan Li, and Bo Han. Learning to augment distributions for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 36, 2024.

[64] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[65] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[66] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in neural information processing systems*, 33:20685–20696, 2020.

[67] 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.

[68] Feng Xue, Zi He, Chuanlong Xie, Falong Tan, and Zhenguo Li. Boosting out-of-distribution detection with multiple pre-trained models. *arXiv preprint arXiv:2212.12720*, 2022.

[69] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

[70] Le Yang, Yizeng Han, Xi Chen, Shiji Song, Jifeng Dai, and Gao Huang. Resolution adaptive networks for efficient inference. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2366–2375, 2020.

[71] 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.

[72] Ji Zhang, Lianli Gao, Bingguang Hao, Hao Huang, Jingkuan Song, and Hengtao Shen. From global to local: Multi-scale out-of-distribution detection. *IEEE Transactions on Image Processing*, 2023.

[73] 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.

[74] 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, 2018.

[75] Ev Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13994–14003, 2020.