Test-Time Linear Out-of-Distribution Detection

Ke Fan[1*], Tong Liu[2*], Xingyu Qiu[1], Yikai Wang[1], Lian Huai[2†], Zeyu Shangguan[2],
Shuang Gou[2], Fengjian Liu[2], Yuqian Fu[1], Yanwei Fu[1], Xingqun Jiang[2]
[1]Fudan University [2]BOE Technology Group
kfan21@m.fudan.edu.cn, {xyqiu20, yikaiwang19, fuyq20, yanweifu}@fudan.edu.cn
{liutongcto, huailian, shangguanzeyu, goushuang, liufengjian, jiangxingqun}@boe.com.cn

Abstract

Out-of-Distribution (OOD) detection aims to address the excessive confidence prediction by neural networks by triggering an alert when the input sample deviates significantly from the training distribution (in-distribution), indicating that the output may not be reliable. Current OOD detection approaches explore all kinds of cues to identify OOD data, such as finding irregular patterns in the feature space, logit space, gradient space, or the raw image space. Surprisingly, we observe a linear trend between the OOD score produced by current OOD detection algorithms and the network features on several datasets. We conduct a thorough investigation, theoretically and empirically, to analyze and understand the meaning of such a linear trend in OOD detection. This paper proposes a Robust Test-time Linear method (RTL) to utilize such linear trends like a 'free lunch' when we have a batch of data to perform OOD detection. By using a simple linear regression as a test time adaptation, we can make a more precise OOD prediction. We further propose an online variant of the proposed method, which achieves promising performance and is more practical for real applications. Theoretical analysis is given to prove the effectiveness of our methods. Extensive experiments on several OOD datasets show the efficacy of RTL for OOD detection tasks, significantly improving the results of base OOD detectors. Project will be available at https://github.com/kfan21/RTL.

1. Introduction

Deep neural networks are renowned for their exceptional performance on image recognition tasks [4, 7, 13, 29].

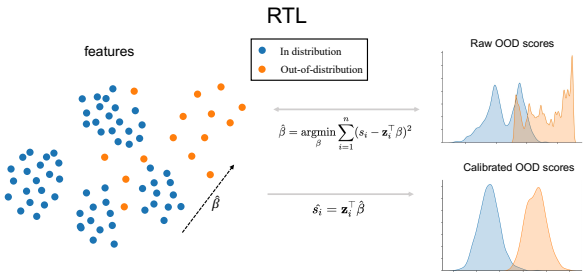

Figure 1. Illustration of our RTL. Blue/orange points denote in/out-of-distribution features. The probability density plot of the OOD score distributions of in and out-of-distribution samples are different. We fit a linear regression between the OOD scores and features as a robust test-time learning method. Then we calibrate the OOD scores to get better OOD predictions.

Nevertheless, the reliability of model predictions is contingent on the input data conforming to the distribution of the training dataset. In situations where the input data significantly deviates from the training data, the neural networks tend to make excessively overconfident predictions [25, 32], thereby impairing the dependability of the deep model in real-world applications. To address this issue, several techniques [8, 11, 14, 18, 19, 21, 27, 31] have been proposed to enable trained models to identify irregular out-of-distribution input data and abstain from generating predictions.

Generally, the task of detecting out-of-distribution (OOD) data typically involves discerning a mixture of test data comprising both in-distribution and data from a distinct dataset [8, 11, 17–19]. Currently, sophisticated algorithms, denoted as f, are predominantly used to derive an (unnormalized) OOD score, $s \in \mathbb{R}$, that indicates the likelihood of the data being OOD. Subsequently, a threshold is applied to the score. To evaluate the efficacy of OOD detection algorithms, all test data must be gathered to determine the threshold and compute false positive rate at 95% recall (FPR95) and area under the receiver operating characteristic curve (AUROC).

The intuition of our method is illustrated in Fig. 1. Al-

*Ke Fan and Tong Liu contribute equally.
†Lian Huai is the corresponding author.
Yanwei Fu is with School of Data Science, Fudan University, Shanghai Key Lab of Intelligent Information Processing, and Fudan ISTBI–ZJNU Algorithm Centre for Brain-inspired Intelligence, Zhejiang Normal University, Jinhua, China.

though OOD detection algorithms typically rely on complex calculations and sophisticated models, we have discovered a surprising linear relationship between the estimated OOD score s and the network feature $\boldsymbol{x}$ before the final fully-connected layers, regardless of the selection of the algorithm f. To establish the validity of this trend, we conducted experiments in Sec. 3.2 using various OOD detection algorithms and benchmark datasets, including iNaturalist and SUN. We found that the linear trend holds across a wide range of scenarios and datasets, indicating its broad applicability. To further confirm the significance of this relationship, we conducted canonical-correlation analysis to measure the strength of the linear correlation between the features and OOD scores. Our analysis shows a strong correlation coefficient, providing the evidence of the relationship's significance. Overall, our findings provide a valuable insight into the nature of OOD detection algorithms and may inform the development of more efficient and accurate models for this task.

We present a novel approach, Robust Test-Time Linear method (RTL) for out-of-distribution detection, to explore the influence of linear trends on OOD benchmarks and enhance OOD detection performance. Our RTL utilizes the linear trend as a "free lunch" to rectify the OOD score directly from input features, with the OOD score initialized using a base OOD detection algorithm. Our experiments demonstrate that this approach significantly improves the performance of various OOD detection algorithms, including those that were originally weak in detecting OOD samples. Essentially, RTL requires a batch of testing data for OOD detection, classifying it as test-time adaptation. However, this setting does not diminish the advantages of RTL over previous OOD methods, as calculating FPR95 and AUROC also demands a batch of testing data for computation. Critically, RTL can be easily implemented as an online version, making it more practical for real-world applications.

Furthermore, To address potential issues of errors introduced by the base OOD detection algorithm, we propose an improved version of our method, RTL++, which reduces the impact of these errors. We provide theoretical analysis to reveal the insights of our RTL improving the base OOD methods. We show that with a moderate error strength, the wrong prediction of these base OOD detectors can be rectified. Empirically, we conduct extensive experiments to demonstrate the effectiveness of RTL across different datasets and base OOD detection algorithms. In summary, RTL is a valuable addition to OOD detection methods, offering a robust and practical solution to enhance detection performance.

To sum up, we identify an important property of OOD benchmarks and propose a novel approach that effectively leverages this property. Through extensive experiments on various datasets and base OOD detection algorithms, we demonstrate the effectiveness of our approach. Specifically, we make the following **contributions**:

- We investigate the presence of a linear trend in current OOD detection benchmarks, where the estimated OOD score is linearly related to the network features before the final fully-connected layer, independent of the OOD detection algorithm employed.
- We propose a straightforward and effective test-time adaptation approach – RTL, that utilizes this linear trend. To improve its robustness, we introduce several variants of RTL, including RTL++, and an online version of RTL for improved practicality.
- We provide theoretical analysis that illuminates the benefits of utilizing the linear trend for OOD detection in test time, providing insights that can inform future research in the field.

2. Related Work

Out-Of-Distribution Detection. In [8], a method for detecting out-of-distribution (OOD) examples using the maximum softmax probability (MSP) of pre-trained neural networks was proposed, analogous to the detection of misclassified examples. ODIN [19] improved MSP using input preprocessing inspired by the idea of adversarial examples [6]. Lee *et al*. [18] introduced a Mahalanobis distance-based method for OOD detection, which achieved better results than MSP but required hyperparameter optimization on a validation or in-distribution set. To address this issue, Liu *et al*. [21] proposed an energy-based score as a parameter-free OOD detector. Further, Huang *et al*. [14] introduced GradNorm, which uses the gradient information to minimize the KL discrepancy between the predicted posterior and the uniform distribution. These *post hoc* OOD detection methods [41] can be applied to any pre-trained classification network trained with cross entropy loss. In contrast, Outlier Exposure (OE) methods [2, 9, 22] directly train an OOD detector using an auxiliary OOD dataset.

Test-Time Adaptation (TTA). TTA is a method introduced in [15, 30, 34] to address the distributional shift problem between training and testing data that can cause deep models to perform poorly on test data from an unseen distribution. During testing, TTA adapts the trained model to the novel data by updating its parameters with a mini-batch or full unlabeled test data. Batch-norm parameters/statistics are updated to fit the testing data [12, 34], or the prediction inconsistency [44] between different data augmentations of a single data point is minimized. TTA is related to but different from Transductive learning (TL), where models are trained jointly on both the train and test data (without testing labels). In contrast, TTA updates the model with the unlabeled test data at test time, as in [34]. Conceptually, our approach follows the TTA setup since we have access to the mini-batch or full test data. However, we differ in our focus on using the inferred OOD signal at test time and

the linear relationship between features and OOD scores to improve OOD detection, rather than adapting the model to the test data. This distinguishes our approach from TTA.

Outlier Detection. Outlier detection [26] is a pressing problem as sampled data is often contaminated with outliers. OOD detectors usually process a single test sample at a time, whereas outlier detectors assume accessibility to all the test samples. However, in the test-time adaptation setup for evaluating OOD detection, the two settings become similar. Therefore, we also compare our algorithms with some outlier detection methods, such as Local Outlier Factor [1], which identifies outliers as samples with substantially lower density than their neighbors, and Isolation Forest [20], which employs decision trees to detect outliers.

3. Methodology

Problem Setup. In image classification tasks, one typically learns a mapping $f : \mathcal{X} \to \mathcal{Y}$ from image space $\mathcal{X} \subseteq \mathbb{R}^m$ to label space $\mathcal{Y} = \{1, 2, \cdots, C\}$ with a given training set $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$. Then at inference, the predicted label can be achieved according to the maximum score $\hat{y} = \operatorname{argmax}_j f_j(\boldsymbol{x})$.

In general, OOD detection involves using continuous OOD scores to identify instances not belonging to these labels by setting a threshold below which they are rejected. For this purpose, we use an OOD score estimator $S(\boldsymbol{x})$ for each $\boldsymbol{x}$ and manually set a threshold γ.

$$g(\boldsymbol{x}) = \begin{cases} \text{in}, & \text{if } S(\boldsymbol{x}) \geq \gamma, \\ \text{out}, & \text{if } S(\boldsymbol{x}) < \gamma. \end{cases} \tag{1}$$

This paper addresses OOD detection for a group of data, either with full access to all test data like TTA, or with a batch of test data, similar to online TTA. To this end, we propose a Robust Test-time Linear method (RTL) that leverages OOD scores generated by a fixed pretrained network with full or batch test data. Importantly, our approach does not require access to the training set and incurs low learning cost. In Table 1, we compare our TTA OOD detection method with others. OOD methods can be broadly classified as either *post hoc* or *outlier exposure*. Post hoc OOD can be applied to any network at little extra cost, while outlier exposure requires both in-distribution and OOD data to train a network. During test time, both methods are applied to samples individually.

Our TTA OOD method only requires additional access to batches of test data in addition to post hoc OOD. This is a reasonable assumption, as in- and out-of-distribution data may come simultaneously. Furthermore, unlike outlier exposure methods, our TTA OOD method does not require retraining the entire network, which can be prohibitively expensive, especially when collecting in-distribution data is difficult or the amount of data is large.

	Access to train data	Access to test dataset	Cost
Post hoc OOD	No	No	Low
Outlier Exposure	Yes	No	High
Outlier Detection	Yes	Yes	High
TTA OOD	No	Yes	Low

Table 1. The difference between several settings.

In addition, we examine outlier detection methods, which detect samples that deviate significantly from the majority by observing all samples. However, this approach is expensive and requires access to both the training and test datasets. Our TTA OOD method, on the other hand, only needs access to the test set. Furthermore, when OOD data constitutes a significant proportion of the dataset, outlier detection methods may fail due to assumptions being violated.

3.1. Baselines

Here we briefly review several recent **OOD score** methods before presenting our contributions in next sections.

MSP. It [8] is defined as OOD score given by a trained network $f_j(\cdot)$ (for class label j).

$$S_{\text{MSP}}(\boldsymbol{x}) := \frac{\exp\left(f_{\hat{y}}(\boldsymbol{x})/T\right)}{\sum_{j=1}^{C} \exp\left(f_j(\boldsymbol{x})/T\right)}, \tag{2}$$

where T is the temperature coefficient to make the softmax prediction sharp, while Vanilla MSP [8] did not include temperature scaling.

Energy. Energy-based model (EBM) [16] aims to find a suitable energy function $E(x, y)$ defined on $\mathcal{X} \times \mathcal{Y}$, and model the posterior probability by the Gibbs distribution

$$p(y \mid \boldsymbol{x}) = \frac{\exp(-E(\boldsymbol{x}, y)/T)}{\exp(-E(\boldsymbol{x})/T)}. \tag{3}$$

The Helmholtz free energy $E(\boldsymbol{x})$ of a given data point $\boldsymbol{x}$ can be expressed as the negative of the log partition function

$$E(\boldsymbol{x}) = -T \cdot \log \int_{y'} \exp(-E(\boldsymbol{x}, y')/T). \tag{4}$$

The Helmholtz free energy for $\boldsymbol{x} \sim P_{in}$ pushes down during the training process, and therefore can serve as an alternative metric for OOD detection [21]. Direct connecting Helmholtz free energy with softmax probability, we get another baseline OOD detector,

$$S_{energy}(\boldsymbol{x}) := -E(\boldsymbol{x}) = T \cdot \log \sum_{i}^{C} \exp(f_i(\boldsymbol{x})/T) \tag{5}$$

KL. Samples from the in-distribution are expected to have low prediction entropy, whereas OOD samples, which are not well-defined in the training label space, should have

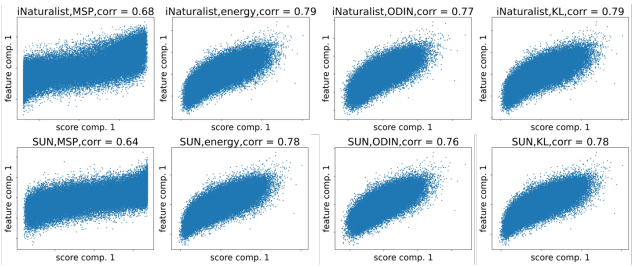

Figure 2. Visualization of Canonical-Correlation Analysis of ImageNet and two OOD datasets' features and OOD scores.

high predictive entropy. To address this, [17] proposed using the Kullback-Leibler (KL) divergence between the softmax output and a uniform distribution to improve OOD scoring.

$$S_{\mathrm{KL}}(\boldsymbol{x}) := D_{\mathrm{KL}}\left(\mathbf{u} \| \mathrm{SoftMax}(f(\boldsymbol{x}))\right) \tag{6}$$

ODIN. It [19] was inspired by adversarial attacks [6] and found that including adversarially perturbed inputs during training improves the final OOD scoring. ODIN generates a perturbed image by first taking an input image $\boldsymbol{x}$ and its predicted softmax probability,

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \varepsilon \cdot \mathrm{sign}\left(-\nabla_{\boldsymbol{x}} \log S_{\mathrm{MSP}}(\boldsymbol{x})\right), \tag{7}$$

Then perturbed image $\tilde{\boldsymbol{x}}$ produces OOD score as,

$$S_{\mathrm{ODIN}}(\boldsymbol{x}) = S_{\mathrm{MSP}}(\tilde{\boldsymbol{x}}). \tag{8}$$

3.2. Robust Test-Time Linear Method

Discussion of the linear relationship We will demonstrate a linear relationship between the extracted features and inferred scores during test time. To test whether there is a correlation, we use canonical-correlation analysis [10] between the features and OOD scores. Specifically, given an in-distribution dataset and a distinct dataset for OOD data, we empirically observe that a simple linear relationship exists between the features extracted from the neural network and the inferred OOD scores by those base OOD detection algorithms, as shown in Fig. 2. Scatter plots between the first pair of canonical components show that there is a rough linear relation between the features and scores. Therefore, although those OOD scores are from different scoring algorithms, the input features and OOD scores are well fitted with a linear regression.

We think that the above four OOD scores are likely to be sub-optimal because they only deal with a single sample and ignore the interaction between samples. In this paper, we show that with a mini-batch (full) test data in hand, we can improve the baseline OOD methods significantly. Despite the presence of some inaccurately inferred OOD scores due to model over-confidence, training a linear regression at test time can rectify these scores.

Mathematical Formulation Let us denote the (unnormalized) OOD score of an input image as $s := S(\boldsymbol{x})$. Due to the linear relationship between features and OOD scores, we assume a linear relation between the OOD score s and the input feature $\mathbf{z}$ extracted by the trained model:

$$s = \mathbf{z}^\top \beta + \varepsilon. \tag{9}$$

Where $\mathbf{z}^\top \beta$ determines whether the image $\boldsymbol{x}$ is in or out-of-distribution samples while ε is the $error$ introduced by OOD detectors. We aim to estimate the β from the feature-score pair $(\mathbf{z}_i, s_i)$; hence we can get a more precise estimation of $\mathbf{z}_i^\top \beta$.

We propose two test-time linear training methods that differ in the type of error they account for. The first for instances that are scored with moderate strength of errors and the second for samples with too much error to train a linear model directly.

RTL (**R**obust **T**est-Time **L**inear Method). When linear relation is recognizable, a simple linear regression model is sufficient to estimate the β value such that:

$$\hat{\beta} = \underset{\beta}{\mathrm{argmin}} \sum_{i=1}^{n} (s_i - \mathbf{z}_i^\top \beta)^2, \tag{10}$$

which yields the closed-form solution

$$\hat{\beta} = (\mathbf{Z}^\top \mathbf{Z})^\dagger \mathbf{Z}^\top \mathbf{S}, \tag{11}$$

where $\mathbf{Z}$ and $\mathbf{S}$ are the stack of $\mathbf{z}_i$ and s_i by rows and $(\cdot)^\dagger$ refer to Moore-Penrose inverse. With this estimator, we could directly provide our OOD estimator for instance i as:

$$\hat{s}_{\mathrm{ours}} = \mathbf{z}_i^\top \hat{\beta}. \tag{12}$$

RTL++. We further consider the challenging case that OOD scoring method does not work good enough. And this results in large errors of inferred OOD scores. Thus, we can inject the outlier detection concept to improve our RTL by removing those data scores before estimating β.

Formally, we present an improved Robust Test-Time Linear Method (RTL++). Specifically, we introduce an explicit data-dependent variable γ_i to model this small amount of large error, such that

$$s_i = \mathbf{z}_i^\top \beta + \gamma_i + \varepsilon_i. \tag{13}$$

The $|\gamma|$ will be very sparse since the just a few samples will have such huge error, while $|\gamma_i|$ will be relatively large when it is not equal to 0. The γ is called incidental parameter, originated from statistics [24] and have applications in many topics [5, 28, 35–38]. We would like to find samples with zero γ_i to fit β. Based on this intuition, we design the following optimization problem:

$$\min_{\beta,\gamma} \sum_{i=1}^{n} \left[\frac{1}{2}(s_i - \mathbf{z}_i^\top \beta - \gamma_i)^2 + \lambda|\gamma_i|\right]. \tag{14}$$

Algorithm 1: Subser selection of RTL++

Input: features $\mathbf{z}_i$ and OOD score s_i, $1 \leq i \leq n$,
Normalize $\mathbf{z}_i$ to unit Euclidean norm
Apply dimensionality reduction on $\mathbf{z}_i$ to $d \ll n$
Stack $\mathbf{z}_i$ and s_i by rows to $\mathbf{Z}$ and $\mathbf{S}$
Calculate projection $\tilde{\mathbf{Z}} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^\dagger\mathbf{Z}^\top$ and $\tilde{\mathbf{S}} = \tilde{\mathbf{Z}}\mathbf{S}$
Solving Lasso $\hat{\gamma} = \text{argmin}_\gamma \frac{1}{2}\|\tilde{\mathbf{S}} - \tilde{\mathbf{Z}}\gamma\|_2^2 + \lambda\|\gamma\|_1$
Select a subset $\hat{\mathbf{Z}}$ with the lowest $p\%$ of $|\hat{\gamma}_i|$.
return $\hat{\mathbf{Z}}$

The ℓ_1 penalty imposed on γ encourages sparse solution. When all γ_i are resolved, we can directly get the closed-form estimation of β as $\hat{\beta} = (\mathbf{Z}^\top\mathbf{Z})^\dagger\mathbf{Z}^\top(\mathbf{S} - \gamma)$. Substituting it into the objective and further defining that $\tilde{\mathbf{Z}} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^\dagger\mathbf{Z}^\top$ and $\tilde{\mathbf{S}} = \tilde{\mathbf{Z}}\mathbf{S}$, we can simplify the objective as

$$\min_\gamma \frac{1}{2}\|\tilde{\mathbf{S}} - \tilde{\mathbf{Z}}\gamma\|_2^2 + \lambda\|\gamma\|_1, \tag{15}$$

which is a standard LASSO problem for γ. We can then take a proper λ to solve γ, as well as the most reliable subset to estimate β as

$$\hat{\beta} = (\mathbf{Z}_{\text{sub}}^\top\mathbf{Z}_{\text{sub}})^\dagger\mathbf{Z}_{\text{sub}}^\top\mathbf{S}_{\text{sub}}, \tag{16}$$

where $(\cdot)_{sub}$ means select a subset of rows to form a new matrix, and we specifically wrote down the selection process in Algorithm 1.

3.3. Theoretical Analysis

This section will explain how our RTL improves the base OOD detectors. The core of our analysis is that when the error strength of base OOD detector is moderate, the wrong prediction can be rectified. We first regard OOD detection as a ranking problem. Suppose there exist a ground-truth OOD score $s = \mathbf{z}^\top\beta$ perfectly rank in and out-of-distribution samples $\mathbf{z}_{in}^\top\beta > \mathbf{z}_{out}^\top\beta$. If the error distorted the ranking

$$\mathbf{z}_{in}^\top\beta + \varepsilon_{in} < \mathbf{z}_{out}^\top\beta + \varepsilon_{out}, \tag{17}$$

it means OOD detectors make mistakes. Our method is based on this insight: the errors of scores can be reduced to a neglected level for ranking when we have enough samples. Although $\mathbf{z}^\top\hat{\beta}$ can be different from $\mathbf{z}^\top\beta$, it will be able to recover the ranking of ID/OOD samples.

Let $s_i, \mathbf{z}_i$ and ϵ_i denote the predicted OOD score, extracted feature and prediction error of the i^{th} example, we have $s_i = \mathbf{z}_i^\top\beta + \epsilon_i$, where ϵ_i are independent and identically distributed (i.i.d.) among different samples. We further assume that:

(C1: Ground-Truth Ranking) The ground-truth score can perfectly rank the ID and OOD samples with a margin $t > 0$, which means $\mathbf{z}_{in}^\top\beta > \mathbf{z}_{out}^\top\beta + t$ for all samples.
(C2: Sub-Gaussian Error) The error ϵ follow a sub-Gaussian distribution SubGau$(0, \nu^2)$ with zero mean and parameters ν^2, where ν is not too larege.

Note that C2 is a weak assumption for detectors. Denote $\hat{\beta}$ as the fitted β of linear regression. We have

Theorem 1 (Ranking Recovery of RTL)**.** *Denote r as the rank of the feature matrix $\mathbf{Z}$ and n as the sample number. If the ratio of the margin to error strength t/ν follow:*

$$t/\nu > min\{2\sqrt{2log(n/\delta)}, 2\sqrt{2r\log(2r/\delta)}\},$$

with at least probability $1-\delta$, $\mathbf{z}_i^\top\hat{\beta}$ rank all ID samples over OOD samples.

Please refer to our supplementary for the full proof and extensive experiments. We conclude that if the error strength ν is kept at a moderate level, even if the base OOD detector could produce predictions that deviate from the ground truth, our RTL can reduce the error for ranking and perfectly detect in and out-of-distribution samples.

It is worth noting that the second part of the bound $t/\nu > 2\sqrt{2r\log(2r/\delta)}$ is irrelevant to the sample numbers n. Since r is always bounded by the dimension of feature space d, this bound is effective when $n \to \infty$. Thus no matter how many samples we get, the model can perfectly rectify OOD scores with high probability. Since $2\sqrt{2r\log(2r/\delta)}$ is a monotone increasing function, the smaller the rank r, the more efficient our RTL. Note that most deep learning models' features lie on a low-dimensional manifold of feature spaces or can be determined by a low-dimensional manifold, we can expect $r \ll d$ for most networks. For example, in Fig. 4, the ImageNet and iNaturalist features can roughly be determined by two-dimensional PCA features.

For RTL++ cases, the model assume another error distribution

$$s_i = \mathbf{z}_i^\top\beta + \gamma_i + \varepsilon_i, \tag{18}$$

due to the sparsity and relatively large scale of γ_i, we can regard $\gamma_i + \varepsilon_i$ follows a mixture of two Sub-Gaussian distribution

$$\gamma_i + \varepsilon_i \sim \pi_1\text{SubGau}(0, \nu_1^2) + \pi_2\text{SubGau}(0, \nu_2^2), \tag{19}$$

where $\pi_1 + \pi_2 = 1$ and π_2 is very small. The two Sub-Gaussian errors have rather different scales $\nu_2 \gg \nu_1$. Direct use of RTL will be difficult since ν_2 determines the error strength due to the following lemma:

Lemma 2. *The mixture of two Sub-Gaussian of SubGau$(0, a^2)$ and SubGau$(0, b^2)$ is still a Guassian Mixture with parameter $max\{a^2, b^2\}$*

Since RTL++ first filters those samples with large errors and then applies linear regression, the analysis of RTL can be applied.

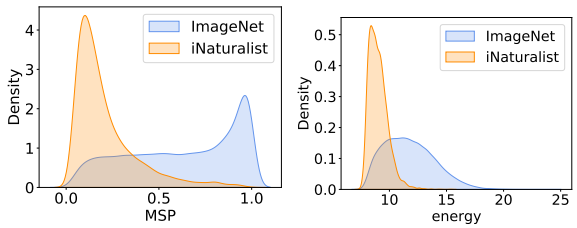

Figure 3. Kernel density estimate plot of the in- and out-of-distribution samples.

3.4. Online OOD Detection

We propose a methodology that assumes the availability of the full set of test instances. However, we acknowledge that in practice, accessing the entire test set may not be feasible. Therefore, we also provide an online version, inspired by previous test-time adaptation works [15, 30, 34].

In the online version, test data arrives in a stream, batch by batch. We derive a batch-wise version of our RTL, which updates the linear model iteratively with the current and past mini-batch data. To estimate β online, we use the wisdom of processing two mini-batch data pairs $(\mathbf{Z}_1, \mathbf{S}_1)$ and $(\mathbf{Z}_2, \mathbf{S}_2)$. We then obtain two block matrices,

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix}, S = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix}.$$

Recall that $\hat{\beta} = (\mathbf{Z}^\top \mathbf{Z})^\dagger \mathbf{Z}^\top \mathbf{S}$, then we can get

$$\mathbf{Z}^\top \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_1^\top & \mathbf{Z}_2^\top \end{bmatrix} \begin{bmatrix} \mathbf{Z}_1 \\ \mathbf{Z}_2 \end{bmatrix} = \mathbf{Z}_1^\top \mathbf{Z}_1 + \mathbf{Z}_2^\top \mathbf{Z}_2, \quad (20)$$

$$\mathbf{Z}^\top \mathbf{S} = \begin{bmatrix} \mathbf{Z}_1^\top & \mathbf{Z}_2^\top \end{bmatrix} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \end{bmatrix} = \mathbf{Z}_1^\top \mathbf{S}_1 + \mathbf{Z}_2^\top \mathbf{S}_2. \quad (21)$$

When more mini-batch data pairs $(\mathbf{Z}_i, \mathbf{S}_i)$ come, we thus can update $\mathbf{Z}^\top \mathbf{Z}$ and $\mathbf{Z}^\top \mathbf{S}$.

4. Experiment Results

Experiment Setting. We conduct experiments on both small scale datasets and large scale datasets. For small scale datasets, we conduct experiments on the CIFAR-10 and CIFAR-100 datasets. Although CIFAR datasets contain easy classification images, it is hard for OOD detection due to its low resolution, especially on CIFAR-100. We use a Wide ResNet [43] trained on CIFAR-10 and CIFAR-100. We use the CIFAR test set as in-distribution data and sample 2000 images from six different out-of-distribution datasets including Textures [3], SVHN [23], Places365 [45], LSUN-Crop [42], LSUN-Resize [42], and iSUN [40] following the setting in [21]. For large scale OOD detection, we use the ImageNet-1k benchmark following [14]. We use the validation set of ImageNet-1k as the in-distribution data, which consists of 50000 natural images with 1000 categories. The out-of-distribution data consist of four datasets, iNaturalist [33], SUN [39], Places [45] and Textures [3].

4.1. Visualization

Given an in-distribution dataset, and we sample the OOD data from a specific dataset. We empirically show that the OOD features behave normally in Fig. 4. We visualize the features with PCA. The OOD features scatter normally in the space, like different classes of data from in-distribution data. This is counter-intuitive, since OOD feature do not disperse in the void of the ID sample like common assumption of outliers. What's more, it is worth noting that OOD data may not just account for a small portion of data. In our example, the test including 50000 in-distribution samples and 10000 OOD samples, which make is extremely hard to apply outlier detection methods.

We draw OOD scores of in-distribution and out-of-distribution data using popular baseline OOD detection algorithms as in Fig. 3. This suggests that the OOD score distributions are not disjoint between in-distribution and out-of-distribution data. Hence they cannot produce a perfect OOD detection. However, we can still observe the different tendency for two kinds of data.

4.2. Small-scale OOD Experiments

Due to randomness, we repeat 10 times trials and report the average results. We set temperature $T = 1$ for all detectors. For ODIN, noise level is set to $\epsilon = 0.0024$. We set $p = 80$ in Algorithm 1. Besides those popular base OOD detectors, we also compare our algorithm with GradNorm [14] and three traditional methods: Gaussian Mixture Models (GMM), Local Outlier Factor (LOF) and Isolation Forest (IF).

The results are shown in Tab. 2, where results show that our test time RTL can improve the base OOD detection methods in almost all cases. Moreover, results in Tab. 2 show that our robust method RTL++ can boost the performance further. Our RTL++ outperforms RTL by 7.9% FPR95 and 2.56% AUROC on CIFAR-100 when using MSP as our base scoring function. Consistent improvements can be observed when using other scoring functions, such as Energy, ODIN and KL. We can also see that our RTL and RTL++ outperforms the other transductive meth-

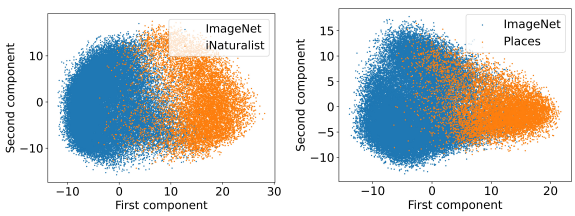

Figure 4. Visualization of features from ImageNet and two OOD datasets. The dimension of features is reduced to two by PCA.

In Dataset	CIFAR-10		CIFAR-100	
Metric	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Softmax	51.37	90.87	80.21	75.67
+RTL	**12.30**	**97.01**	51.63	84.50
+RTL++	13.50	96.42	**43.73**	**87.06**
Energy	32.98	91.88	73.46	79.67
+RTL	18.01	94.96	60.63	81.15
+RTL++	16.14	95.64	58.06	82.43
ODIN	35.77	90.96	74.55	77.23
+RTL	37.94	86.10	51.87	81.52
+RTL++	35.79	87.24	51.73	82.74
KL	32.98	91.88	73.46	79.67
+RTL	18.01	94.96	60.63	81.15
+RTL++	16.12	95.64	58.06	82.43
IF	79.96	62.47	80.91	66.15
LOF	95.81	56.45	98.23	43.32
GMM	87.70	58.28	94.06	69.96
GradNorm	59.84	71.65	86.55	57.56

Table 2. Results on CIFAR. The best and second best results are highlighted by fonts of text bold and underlined, respectively.

ods IF, LOF and GMM and an advanced OOD method GradNorm. As explained in Sec. 4.1, the assumption of outlier detection is violated and IF, LOF and GMM fails. It is worth noting that our algorithm perform even competitive with Outlier Exposure [9] and fine-tuned energy score [21], which utilize an extra outlier dataset to direct train the network to distinguish in- and out-of-distribution data.

4.3. Large-scale OOD Experiments

Google BiT-S models of ResNetV2-101 trained on ImageNet-1k is used as the feature extractor. For MSP, energy score and KL divergence, we set temperature $T = 1$. For ODIN, temperature is set to $T = 1000$, with noise level of $\varepsilon = 0$ since FGSM will not improve the results. We set $p = 95$ for MSP and ODIN, $p = 90$ for energy score and KL divergence in RTL++ in Algorithm 1.

The results are shown in Tab. 3, where every OOD set are test separately and the mean results are calculated. The results show that our algorithms achieve consistent improvement over the four baseline OOD detectors. For MSP, our linear revision improve FPR95 from 76.96% to 45.97%, about the 30.99% of the improvement, and outperform GradNorm by 8.73%. For AUROC, the linear calibrated MSP produce the best result. In most experiments our RTL++ can bring further gain, which proves the efficacy of filtering large error prediction.

4.4. Online Test-Time Adaptation

In the previous experiments, the full test set is accessed during OOD detection, which is a special case of online test-time learning [15, 30, 34], *i.e*. the batch size equals the total number of test data. Here we further check how the performance varies when the batch size gets smaller, *i.e*. the test time data comes in stream. We compare our online RTL with those popular base OOD methods in this online setup in Tab. 4. From the results, we can see that our online RTL still consistently outperforms the base OOD detection methods, only except ODIN on CIFAR-10, demonstrating the efficacy of our online RTL. More importantly, the results of our online RTL are close to the results in Tab. 2, where the full test data set was accessed during model training. These nearly identical results clearly indicates that the effectiveness of our proposed method is not from the access of the full test set, but from the effective test-time learning as those TTA methods [15, 30, 34].

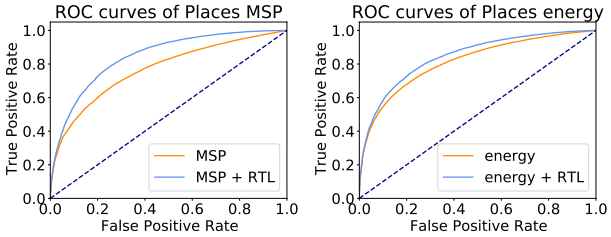

Figure 5. The ROC curves of baseline OOD detector and RTL.

4.5. Ablation Study

ROC curves analysis of RTL We visualize the receiver operating characteristic curve to illustrate the effectiveness of our proposed RTL in Fig. 5. The ROC curves of RTL can wrap up the curves of baseline OOD detectors, providing a better AUROC. What's more, intersection points of a horizontal line with the two curves indicate our RTL give smaller false positive rate when the true positive rates are kept at the same level.

Effect of Subset Selection of RTL++ We plot the performance improvement of our RTL++ as a function of the percentile of chosen subset on CIFAR-100 over two base OOD scoring metrics MSP and ODIN, as shown in Fig. 6. The percentile of chosen data varies from 0.5 to 1, with step 0.05. When percentile is 1, it is equivalent to our RTL. No matter with MSP or ODIN, when we decrease the percentile, the AUROC and FPR95 improve first, which means our RTL++ selects proper subsets for learning the regression, eliminating the samples with large error. When we further decrease the percentile, the performances start getting worse as the training data is too limited to learn a good regressor.

Batch Size Effect of Online RTL We check how the performance of our online RTL varies upon batch size numbers. We run our online RTL with MSP in Tab. 5 by varying the batch size from 32 to *full* on CIFAR OOD benchmarks. From the results we can see that the performance of our online RTL is insensitive to batch size number. The last

OOD Dataset	iNaturalist		SUN		Places		Textures		Average	
Metric	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95	AUROC	FPR95↓	AUROC↑
IF	88.58	61.60	90.12	57.85	93.45	50.24	**54.34**	**87.76**	81.62	64.36
LOF	95.16	51.57	94.89	52.27	93.05	56.37	82.02	65.39	91.28	56.40
GMM	87.90	68.43	89.99	63.29	96.85	52.83	95.37	35.34	92.53	54.97
GradNorm	50.03	90.33	**46.48**	89.03	60.86	84.82	61.42	81.07	54.70	86.31
MSP	63.69	87.59	79.98	78.34	81.44	76.76	82.73	74.45	76.96	79.29
+RTL	21.03	94.98	50.68	87.14	57.22	84.48	58.48	80.24	46.85	86.71
+RTL++	**18.76**	**95.60**	48.40	88.70	**56.72**	**85.32**	59.98	79.91	**45.97**	**87.38**
Energy	64.91	88.48	65.33	85.32	73.02	81.37	80.87	75.79	71.03	82.74
+RTL	45.48	91.04	52.06	88.68	62.68	84.35	69.49	75.39	57.43	84.87
+RTL++	41.57	92.03	49.84	**89.32**	62.37	84.05	70.44	76.52	56.06	85.48
KL	64.91	88.48	65.32	85.31	73.02	81.37	80.87	75.79	71.03	82.74
+RTL	45.48	91.04	52.06	88.68	62.68	84.35	69.49	75.39	57.43	84.87
+RTL++	41.57	92.03	49.84	**89.32**	62.37	84.05	70.44	76.52	56.06	85.48
ODIN	62.69	89.36	71.67	83.92	76.27	80.67	81.31	76.30	72.99	82.56
+RTL	35.27	92.87	51.59	88.40	60.71	84.44	66.72	76.78	53.57	85.62
+RTL++	36.10	92.78	51.87	88.23	61.35	84.28	67.06	76.58	54.09	85.47

Table 3. Results on ImageNet-1k and iNaturalist/SUN/Places/Textures datasets. The best and second-best results are highlighted in bold and underlined font, respectively.

In Dataset	CIFAR-10		CIFAR-100	
Metric Size	FPR95↓	AUROC↑	FPR95↓	AUROC↑
MSP	51.37	90.87	80.21	75.67
+Online RTL	15.06	96.05	55.84	83.31
Energy	32.98	91.88	73.46	79.67
+Online RTL	18.57	94.89	62.00	80.91
ODIN	35.77	90.96	74.55	77.23
+Online RTL	40.30	85.15	54.66	80.24
KL	32.98	91.88	73.46	79.67
+Online RTL	18.57	94.89	62.00	80.91

Table 4. Results on CIFAR under different OOD detector.

In Dataset	CIFAR-10		CIFAR-100	
Batch Size	FPR95↓	AUROC↑	FPR95↓	AUROC↑
Raw Softmax	51.37	90.87	80.21	75.67
32	15.06	96.05	55.84	83.31
64	15.04	96.07	55.77	83.34
128	14.99	96.10	55.64	83.38
256	14.79	96.24	55.29	83.53
512	14.53	96.38	54.79	83.69
1024	14.16	96.52	53.98	83.89
All data	**12.30**	**97.01**	**51.63**	**84.50**

Table 5. Online RTL on CIFAR with different batch size.

row utilize all data for a batch, which is essentially the results of ordinary RTL. The gap between online version and non-online version is small, proving the practicality of our methods.

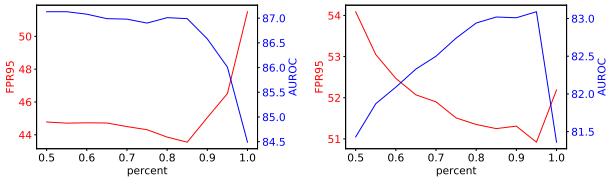

Figure 6. RTL++ results with different percentile of selected subset data on CIFAR-100 with MSP(left) and ODIN(right).

5. Conclusion

In this paper, we present an investigation of the linear trend observed between the OOD score produced by current OOD detection algorithms and the network features on several datasets. Drawing from this observation, we propose a robust test-time linear learning model, RTL, to rectify the original OOD scores. To address cases where the original OOD scores are noisy, we introduce a variant of RTL, named RTL++. Our methods significantly enhance the performance of popular OOD detection methods on various benchmarks. Moreover, we conduct theoretical analysis of the linear trend and provide a sufficient condition to ensure the effectiveness of RTL. Furthermore, We propose an online version of RTL to demonstrate the practicality of our approach, which yields promising results in the online test-time adaptation setup. Based on these results, we suggest that treating OOD detection from test-time adaptation and online learning perspectives is vital and more practical for read-world applications.

References

[1] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000. 3

[2] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 430–445. Springer, 2021. 2

[3] 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. 6

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[5] Yanwei Fu, Timothy M Hospedales, Tao Xiang, Jiechao Xiong, Shaogang Gong, Yizhou Wang, and Yuan Yao. Robust subjective visual property prediction from crowdsourced pairwise labels. *IEEE transactions on pattern analysis and machine intelligence*, 38(3):563–577, 2015. 4

[6] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, 2015. 1

[8] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 1, 2, 3

[9] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 2, 7

[10] Harold Hotelling. Relations between two sets of variates. In *Breakthroughs in statistics*, pages 162–190. Springer, 1992. 4

[11] 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. 1

[12] Xuefeng Hu, Gokhan Uzunbas, Sirius Chen, Rui Wang, Ashish Shah, Ram Nevatia, and Ser-Nam Lim. Mixnorm: Test-time adaptation through online normalization estimation. *arXiv preprint arXiv:2110.11478*, 2021. 2

[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. 1

[14] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *arXiv preprint arXiv:2110.00218*, 2021. 1, 2, 6

[15] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 6, 7

[16] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 3

[17] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017. 1, 4

[18] 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, 2

[19] 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. 1, 2, 4

[20] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008. 3

[21] Weitang Liu, Xiaoyun Wang, John D Owens, and Yixuan Li. Energy-based out-of-distribution detection. *arXiv preprint arXiv:2010.03759*, 2020. 1, 2, 3, 6, 7

[22] Marc Masana, Idoia Ruiz, Joan Serrat, Joost van de Weijer, and Antonio M Lopez. Metric learning for novelty and anomaly detection. *arXiv preprint arXiv:1808.05492*, 2018. 2

[23] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 6

[24] Jerzy Neyman and Elizabeth L Scott. Consistent estimates based on partially consistent observations. *Econometrica: journal of the Econometric Society*, pages 1–32, 1948. 4

[25] 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. 1

[26] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2):1–38, 2021. 3

[27] 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. 1

[28] Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, 106(494):626–639, 2011. 4

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[30] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *In-*

ternational Conference on Machine Learning, pages 9229–9248. PMLR, 2020. 2, 6, 7

[31] Engkarat Techapanurak, Masanori Suganuma, and Takayuki Okatani. Hyperparameter-free out-of-distribution detection using cosine similarity. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1

[32] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[33] 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. 6

[34] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 2, 6, 7

[35] Yikai Wang, Chengming Xu, Chen Liu, Li Zhang, and Yanwei Fu. Instance credibility inference for few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12836–12845, 2020. 4

[36] Yikai Wang, Li Zhang, Yuan Yao, and Yanwei Fu. How to trust unlabeled data? instance credibility inference for few-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6240–6253, 2021.

[37] Yikai Wang, Xinwei Sun, and Yanwei Fu. Scalable penalized regression for noise detection in learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 346–355, 2022.

[38] Yikai Wang, Yanwei Fu, and Xinwei Sun. Knockoffs-spr: Clean sample selection in learning with noisy labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4

[39] 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. 6

[40] 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. 6

[41] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 2

[42] 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. 6

[43] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 6

[44] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *arXiv preprint arXiv:2110.09506*, 2021. 2

[45] 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. 6