READ: Aggregating Reconstruction Error into Out-of-Distribution Detection

Wenyu Jiang, Yuxin Ge, Hao Cheng, Mingcai Chen, Shuai Feng, Chongjun Wang

State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{lygjwy, yuxinge, chengh, chenmc, shuaifeng}@smail.nju.edu.cn, chjwang@nju.edu.cn

Abstract

Detecting out-of-distribution (OOD) samples is crucial to the safe deployment of a classifier in the real world. However, deep neural networks are known to be overconfident for abnormal data. Existing works directly design score function by mining the inconsistency from classifier for in-distribution (ID) and OOD. In this paper, we further complement this inconsistency with reconstruction error, based on the assumption that an autoencoder trained on ID data can not reconstruct OOD as well as ID. We propose a novel method, READ (**R**econstruction **E**rror **A**ggregated **D**etector), to unify inconsistencies from classifier and autoencoder. Specifically, the reconstruction error of raw pixels is transformed to latent space of classifier. We show that the transformed reconstruction error bridges the semantic gap and inherits detection performance from the original. Moreover, we propose an adjustment strategy to alleviate the overconfidence problem of autoencoder according to a fine-grained characterization of OOD data. Under two scenarios of pre-training and re-training, we respectively present two variants of our method, namely READ-MD (**M**ahalanobis **D**istance) only based on pre-trained classifier and READ-ED (**E**uclidean **D**istance) which retrains the classifier. Our methods do not require access to test time OOD data for fine-tuning hyperparameters. Finally, we demonstrate the effectiveness of the proposed methods through extensive comparisons with state-of-the-art OOD detection algorithms. On a CIFAR-10 pre-trained WideResNet, our method reduces the average FPR@95TPR by up to 9.8% compared with previous state-of-the-art.

Introduction

Deep neural networks (DNNs) have attained high accuracy in image classification task (Zagoruyko and Komodakis 2016). However, the classifier often fails silently by providing overconfident prediction for input that belongs to a distribution different from the in-distribution (ID) of training data. Therefore, it is necessary to detect those out-of-distribution (OOD) samples for the deployment of classifier in safety-critical applications, such as autonomous driving and medical diagnosis.

For detecting OOD samples, the baseline method (Hendrycks and Gimpel 2016) utilizes the maximum value of posterior distribution from the pre-trained softmax classifier. They find that ID data tends to have greater prediction probabilities than OOD data. By temperature scaling and input perturbation, ODIN (Liang, Li, and Srikant 2017) improves the baseline method. However, it has been observed that softmax classifier can produce high confidence prediction for inputs far away from the training data (Hendrycks and Gimpel 2016; Nguyen, Yosinski, and Clune 2015). The rationale is that the softmax classifier can have a label-overfitted output space (Lee et al. 2018; Liu et al. 2020). Instead of using the softmax outputs for OOD detection, Maha (Lee et al. 2018) assumes that pre-trained features of test data can be fitted well by a class-conditional Gaussian distribution and defines the confidence score using the Mahalanobis distance with respect to the closest class-conditional distribution in feature spaces. From probabilistic perspective of decomposing confidence, G-ODIN (Hsu et al. 2020) uses a dividend/divisor structure for classifier. Then, the distance of input to the closest class is calculated with penultimate layer output of classifier to detect OOD samples.

The above methods are based on the observation that OOD data should be relatively far away from the ID classes. In this paper, we further complement the discrepancy of distance to the closest class in latent space. Based on the assumption that test data from the distribution same as training data can be better reconstructed than other distributions, we propose a reconstruction error aggregated detector (READ). The extracted representations by autoencoder are enforced to contain important regularities of the ID data. However, OOD inputs are poorly reconstructed from the resulting representations due to the irregular patterns. Our high-level idea is to mine the discrepancy of ID and OOD from classifier and autoencoder. To unify both discrepancies, i.e., the distance to the closest class and reconstruction error, we transform raw pixels reconstruction error to the latent space of classifier. Overall, the transformed reconstruction error exhibits competitive OOD detection performance compared with the raw pixels. Based on the same reconstruction error assumption, Gong et al. (Gong et al. 2019) and Zhang et al. (Zhang et al. 2021) incorporate memory module to autoencoder and directly use raw pixels reconstruction error to detect OOD samples. However, they find that this assumption does not always hold and the autoencoder can reconstruct specific OOD data well with low reconstruction error.

Similar overconfident phenomenon for flow-based deep generative models is reported in (Choi, Jang, and Alemi 2018; Nalisnick et al. 2018). For transformed reconstruction error, we observe that the same overconfidence problem. In order to alleviate this problem, we further propose a fine-grained characterization of OOD based on (Hsu et al. 2020). Then, we introduce a coefficient to adjust transformed reconstruction error according to the data types. Empirical result shows that adjustment coefficient alleviates the overconfidence problem. Under two scenarios of pre-training and retraining, we propose corresponding variants of READ, namely READ-MD (**M**ahalanobis **D**istance) only based on pre-trained classifier and READ-ED (**E**uclidean **D**istance) which retrains the modified classifier.

The complete illustration of our method is presented in Figure 1. Through extensive and comprehensive evaluations on common OOD detection benchmarks, both of our methods, READ-MD and READ-ED, achieve state-of-the-art performance compared with previously best methods under corresponding scenarios. In ablation studies, we also demonstrate the effectiveness of the proposed transformed reconstruction error and adjustment coefficient. Note that the choice of **hyperparameters does not rely on test time OOD data** and **no auxiliary OOD samples are provided at training time.**

Our main contributions are summarized as follows:

- We propose a novel reconstruction error aggregated detector (READ) and its two variants, READ-MD and READ-ED, which combine the distance to the closest class and reconstruction error in the latent space of classifier.
- Against the overconfidence of transformed reconstruction error, we explain and alleviate this problem by a fine-grained characterization of OOD data and an image complexity based adjustment coefficient.
- We conduct comprehensive analysis with experiments under both scenarios to demonstrate the effectiveness of the proposed methods.

Related Work

Out-of-distribution Detection for Discriminative Models. Given pre-trained classifier, a baseline method (Hendrycks and Gimpel 2016) utilizing maximum softmax probability (MSP) is proposed based on the observation that ID samples tend to have greater prediction probabilities than OOD samples. However, the MSP score for OOD input is proven to be arbitrarily high for neural networks with ReLU activation (Hein, Andriushchenko, and Bitterwolf 2019). Liang et al. (Liang, Li, and Srikant 2017) improve the baseline with temperature scaling and input perturbation techniques, and further enlarge the gap between ID and OOD data. Instead of deriving score function from label-overfitted output space, Lee et al. (Lee et al. 2018) and Sastry et al. (Sastry and Oore 2020) design confidence score in feature spaces of the pre-trained classifier. Liu et al. (Liu et al. 2020) propose energy score which can be easily derived from the logit output of the pre-trained classifier and demonstrate superiority over softmax score both empirically and theoretically. Sun et al. (Sun, Guo, and Li 2021) show that a simple activation rectification strategy termed ReAct can significantly improve OOD detection performance. Recent work by Huang et al. (Huang, Geng, and Li 2021) proposes a score function named GradNorm from the gradient space. GradNorm utilizes the vector norm of gradients, backpropogated from the KL divergence between the softmax output and a uniform probability distribution.

Loosening the restriction on retraining, G-ODIN (Hsu et al. 2020) modifies the classifying head with a dividend/-divisor structure for decomposing confidence of predicted class probabilities. Moreover, a modified input perturbation strategy is proposed to remove the unrealistic requirement of previous methods (Liang, Li, and Srikant 2017; Lee et al. 2018) that the choice of hyperparameters depends on test time OOD data. In (Tack et al. 2020; Sehwag, Chiang, and Mittal 2021), self-supervised learning is used to learn better visual representations for OOD detection. In our work, we further complement the discrepancy of ID and OOD from discriminative models with reconstruction error.

Out-of-distribution Detection for Generative Models. There are several works that detect OOD samples with generative models. The input data is defined as OOD if it lies in the low-density regions. However, as shown in (Choi, Jang, and Alemi 2018; Nalisnick et al. 2018), flow-based generative models (Kingma and Welling 2013; Van den Oord et al. 2016; Rezende, Mohamed, and Wierstra 2014) can assign a high likelihood to OOD data. This problem is addressed by considering a likelihood ratio (Ren et al. 2019), taking the input complexity into account (Serrà et al. 2019), and likelihood regret (Xiao, Yan, and Amit 2020). For autoencoder, the specific OOD inputs can be reconstructed well demonstrated by (Denouden et al. 2018; Gong et al. 2019; Zhang et al. 2021). In contrast, we transform the reconstruction error to latent space of classifier as a supplement and propose an adjustment coefficient to alleviate the overconfidence problem.

Characterization of Out-of-distribution Data. According to (Hsu et al. 2020), OOD data can be described from covariate shift and semantic shift perspectives. Works in OOD detection task focus on detection of semantic shift OOD inputs. We further characterize the semantic shift OOD data based on image complexity (Lin, Roy, and Li 2021; Serrà et al. 2019) for explaining and alleviating the varying detection performance of autoencoder. Some works propose a fine-grained characteization of covariate shift OOD data (Hendrycks and Dietterich 2019; Ovadia et al. 2019) for evaluation of model, including corruption-shift for robustness and domain-shift for domain generalization performance. Note that our work is dedicated to finding new concepts at inference time.

Method

Our method is illustrated in Figure 1. In this section, we first formalize the out-of-distribution detection problem. Secondly, we present the overall concept of our algorithm. Then, we detail the actual training process. To avoid confusion, the

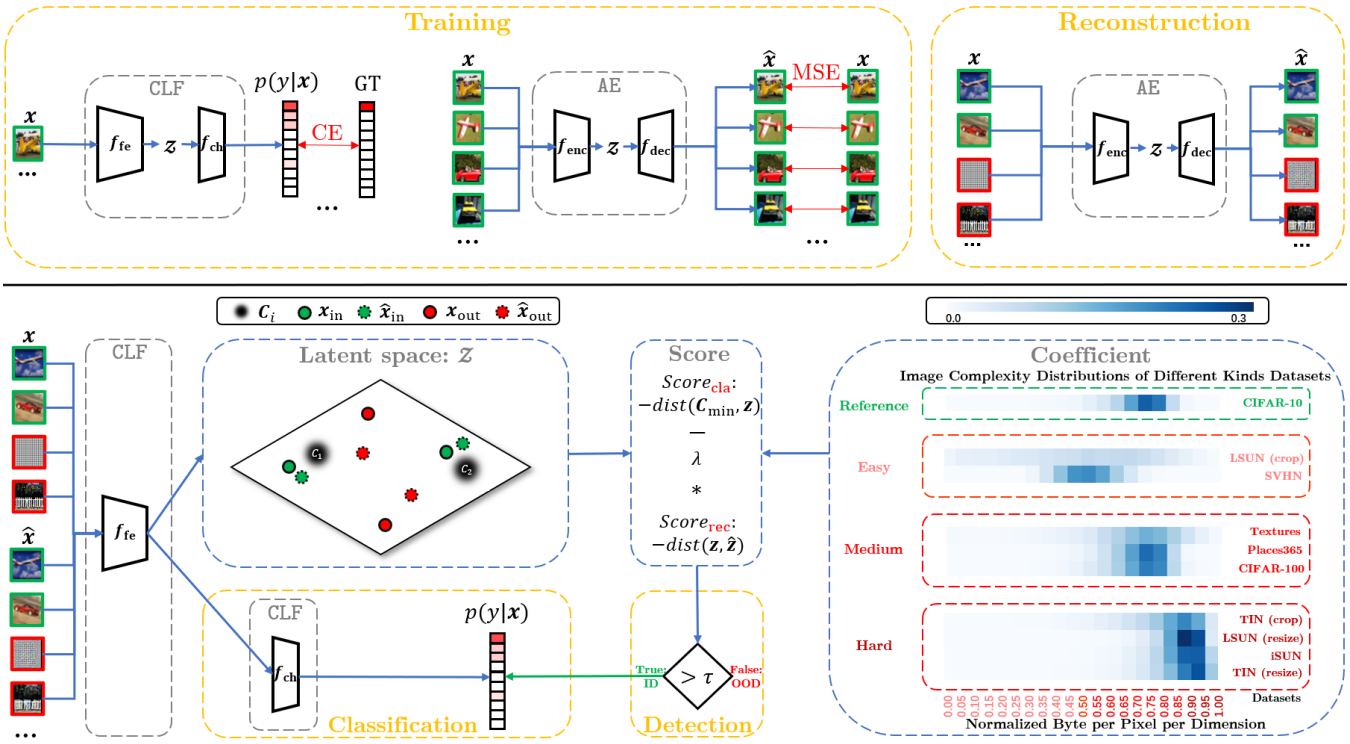

Figure 1: Illustration of the proposed reconstruction error aggregated detector (READ). *Top*: Architecture of the proposed detector and preparatory phase (training and reconstruction). *Bottom*: Overview of the OOD detection and classification procedure at detection phase (detection and classification).

score function is illustrated separately under both scenarios. We explain the reconstruction error adjustment coefficient based on image complexity. Finally, we present the OOD detection and classification procedure at inference time.

Problem Statement

In this paper, we consider the OOD detection problem under setting of multi-category image classification. Let $\mathcal{X} = \mathbb{R}^d$ denote the input space and $\mathcal{Y} = \{1, ..., K\}$ denote the corresponding label space. We have access to the labeled training set $\mathcal{D}_{\text{in}}^{\text{train}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, drawn *i.i.d* from the joint data distribution $\mathbb{P}_{\mathcal{X}\times\mathcal{Y}}$. Let $f_\theta : \mathcal{X} \mapsto \mathbb{R}^{|\mathcal{Y}|}$ denote a neural network for the classification task, which predicts the label of an input sample. Furthermore, we denote the marginal probability distribution on $\mathcal{X}$ by $\mathbb{P}_{\mathcal{X}}^{\text{in}}$, which represents the distribution of in-distribution data. At inference time, the classifier f will encounter a different distribution $\mathbb{P}_{\mathcal{X}^{\text{out}}}$ over $\mathcal{X}$ of out-of-distribution data. Out-of-distribution detection aims to design a binary function estimator,

$$g(x) = \begin{cases} 1, & \text{if } \boldsymbol{x} \sim \mathbb{P}_{\mathcal{X}}^{\text{in}} \\ 0, & \text{if } \boldsymbol{x} \sim \mathbb{P}_{\mathcal{X}}^{\text{out}} \end{cases}$$

that classifies whether a test-time sample $\boldsymbol{x} \in \mathcal{X}$ is generated from $\mathbb{P}_{\mathcal{X}}^{\text{in}}$ or $\mathbb{P}_{\mathcal{X}}^{\text{out}}$. In practice, the $\mathbb{P}_{\mathcal{X}}^{\text{out}}$ is often defined by an irrelevant distribution with non-overlapping labels with regard to in-distribution data. Hence, the classifier f should not predict OOD data.

Overall Concept

Based on the reconstruction error assumption, we introduce autoencoder into out-of-distribution detection. As shown in the Training part of Figure 1, the network architecture of our method consists of two components: (a) a classifier (CLF) containing feature extractor f_{fe} for learning latent representations $\boldsymbol{z}$ with parameters θ_{fe} and classifying head f_{ch} with parameters θ_{ch} which takes $\boldsymbol{z}$ and classify them to known classes. The notation $p(y|\boldsymbol{x})$ denotes the prediction posterior distribution for input $\boldsymbol{x}$. (b) an autoencoder (AE), including encoder f_{enc} to compress high-dimensional data features with parameters θ_{enc} and deconder f_{dec} recreating $\boldsymbol{x}$ denoted by $\hat{\boldsymbol{x}}$ from the latent representation $\boldsymbol{z}$ with parameters θ_{dec}. Different from works (Oza and Patel 2019; Zhang et al. 2021) integrating classifier and autoencoder in one hybrid model simultaneously and utilizing raw pixels reconstruction error as score function, the CLF and AE modules in our method are independent of each other. Furthermore, we transform the reconstruction error to CLF latent space instead of pixels space for further aggregation. For one thing, the transformed reconstruction error bridges the semantic gap, and for another, we empirically show its superior detection performance. Lee at el. (Lee et al. 2018) also remark that OOD samples can be characterized better by latent embeddings of CLF, rather than the "label-overfitted" output space. These take us to transform reconstruction error to the latent space of CLF.

After transforming the reconstruction error, we combine

it with the distance of the input $\boldsymbol{x}$ to the closest category $\mathcal{C}_i$ in the latent space of CLF since the OOD inputs should be relatively far away from the ID classes. We illustrate our idea in latent space part of Figure 1. Obviously, the combination of transformed reconstruction error and distance to the closest class in latent space brings better separability of ID and OOD samples. We proceed with detailing the training procedure, OOD detection and classification procedures.

Training

As described above, our architecture contains two independent components: CLF and AE. In training stage, we need to train CLF to classify ID samples correctly and train AE to reconstruct original inputs. Specifically, our CLF is trained to optimize parameters θ_{fe} and θ_{ch} by minimizing the following cross-entropy loss function:

$$\mathcal{L}_{\text{CLF}} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{in}}^{\text{train}}}[-\log F_y(\boldsymbol{x})] \tag{1}$$

where $F_y(\boldsymbol{x})$ is the softmax output of CLF. For the AE, the parameters θ_{enc} and θ_{dec} are updated. The training mean-squared error loss function is as follows:

$$\mathcal{L}_{\text{AE}} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}_{\text{in}}^{\text{train}}}[\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2^2] \tag{2}$$

where $\hat{\boldsymbol{x}}$ is the reconstruction output of AE and $\mathcal{X}_{\text{in}}^{\text{train}}$ is the ID training data without labels. The complete training procedures are illustrated in Training part of Figure 1.

Transformed Reconstruction Error

To measure the distance of test input $\boldsymbol{x}$ to the closest category $\mathcal{C}_i$ and reconstruction error in latent space $\mathcal{Z}$, we first need to model classes by ID training data. Considering the limitations of CLF retraining in practical problems, we adopt two different class modeling strategies with corresponding distance metrics, namely Mahalanobis distance and Euclidean distance.

Pre-training Scenario. Given the pre-trained CLF without subsequent retraining, we use the same class modeling method as (Lee et al. 2018). We define K class-conditional distributions with a tied covariance $\boldsymbol{\Sigma}$: $p(f_{\text{fe}}(\boldsymbol{x})|y=i) = \mathcal{N}(f_{\text{fe}}(\boldsymbol{x})|\mu_i,\boldsymbol{\Sigma})$, where μ_i is the mean of multivariate Gaussian distribution of class $i\in\{1,...K\}$, assuming that the class-conditional distribution of CLF latent representations follows multivariate Gaussian distribution. Then, the empirical class mean and covariance of training data are computed to estimate the parameters of the class-conditional distribution as follows:

$$\hat{\mu}_i = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}_{\text{in}}^{\text{train}},y=i}[f_{\text{fe}}(\boldsymbol{x})] \tag{3}$$

$$\hat{\boldsymbol{\Sigma}} = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}_{\text{in}}^{\text{train}}}[(f_{\text{fe}}(\boldsymbol{x})-\hat{\mu}_y)(f_{\text{fe}}(\boldsymbol{x})-\hat{\mu}_y)^{\mathsf{T}}] \tag{4}$$

After modeling the ID classes with multivariate Gaussian distributions, we measure distance between test input $\boldsymbol{x}$ and the closest class-conditional distribution by Mahalanobis distance, i.e.,

$$Score_{\text{cla}} = -\min_i (f_{\text{fe}}(\boldsymbol{x})-\hat{\mu}_i)^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}^{-1}(f_{\text{fe}}(\boldsymbol{x})-\hat{\mu}_i)) \tag{5}$$

Accordingly, the definition of reconstruction error between original data $\boldsymbol{x}$ and reconstructed data $\hat{\boldsymbol{x}}$ in latent space of CLF is presented below:

$$Score_{\text{rec}} = -((f_{\text{fe}}(\boldsymbol{x})-f_{\text{fe}}(\hat{\boldsymbol{x}}))^{\mathsf{T}}\hat{\boldsymbol{\Sigma}}^{-1}(f_{\text{fe}}(\boldsymbol{x})-f_{\text{fe}}(\hat{\boldsymbol{x}})) \tag{6}$$

Retraining Scenario. Loosening restriction on the retraining of CLF, Hsu at el. (Hsu et al. 2020) change the original CLF's f_{ch} from fully connected layer to a dividend/divisor structure with a novel perspective of decomposed confidence. Inspired by this, we modify the f_{ch} based on their work as follows:

$$f_{\text{ch}_i}(\boldsymbol{z}) = \frac{h_i(\boldsymbol{z})}{g(\boldsymbol{z})} = \frac{-\|\boldsymbol{z}-\omega_i\|_2^2}{\sigma(BN(\omega_{\text{g}}\boldsymbol{z}+b_{\text{g}}))} \tag{7}$$

where $\boldsymbol{z}$ is the output of f_{fe} and BN denotes the batch normalization layer. In proportion, the ID class centers are fitted by learnable parameters of classifier, i.e., w_i. Moreover, the distance of input $\boldsymbol{x}$ to the closest class center and transformed reconstruction error in latent space of CLF are defined using Euclidean distance as follows:

$$Score_{\text{cla}} = -\min_i(\|\boldsymbol{z}-\omega_i\|_2^2) \tag{8}$$

$$Score_{\text{rec}} = -(\|\boldsymbol{z}-\hat{\boldsymbol{z}}\|_2^2) \tag{9}$$

where $\hat{\boldsymbol{z}}$ is the output of f_{fe} when the input is $\hat{\boldsymbol{x}}$. Note that we do not use auxiliary OOD training data in both scenarios.

Adjustment Coefficient

Although the transformed reconstruction error brings superior discrimination, we observe that the detection performance is inconsistent across various OOD datasets, as shown in Figure 2. Under different metrics, we find the same three distribution patterns of OODs when CIFAR-10 is taken as ID reference. Specifically, the distribution is skewed to the smaller reconstruction error for "easy" OODs which contain simple objects or constant pixels since simpler representations are required for their description. For "medium" OODs which have the covariate same as ID, the distributions are similar, i.e., it is indistinguishable of inputs by reconstruction error. In general, the fact that **AE trained on ID data can reconstruct the "easy" and "medium" OOD data well with low reconstruction error** poses a challenge to OOD detection. Explaining from the multi-category learning process of AE, the diversity of training data increases the difficulty of ID reconstruction compared to the single class. Similarly, the overconfident phenomenon is also reported in (Choi, Jang, and Alemi 2018; Denouden et al. 2018; Gong et al. 2019; Nalisnick et al. 2018; Zhang et al. 2021). Lastly, the reconstruction error distribution of "hard" OODs which contain richer contents and diverse pixels compared to ID is skewed to the right side as expected, and this is consistent with the reconstruction error assumption. The rationale is that the learned representations by AE are enforced to learn important regularities of the ID data to minimize reconstruction errors. Hence, OOD data are poorly reconstructed from the resulting representations.

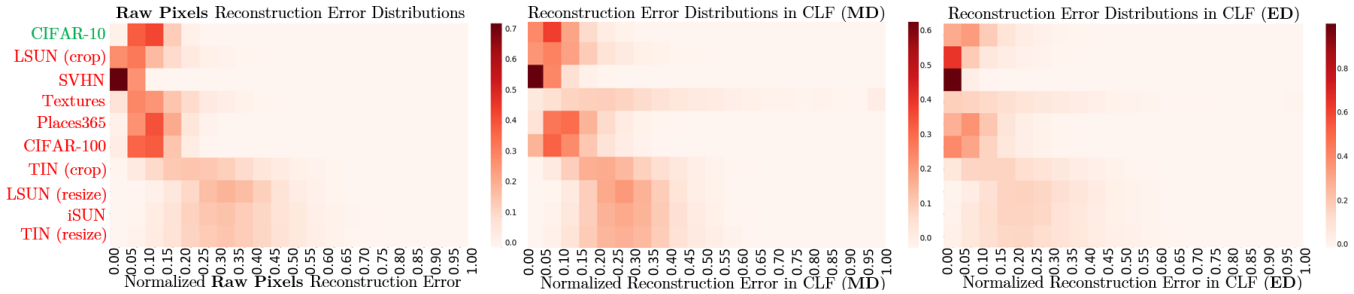

Figure 2: The reconstruction error distributions in different forms (CIFAR-10 as ID).

To sum up, the reconstruction error assumption does not always hold for different kinds of OOD data and this conclusion is applicable to different reconstruction error forms.

In order to alleviate the above issue, we firstly propose a fine-grained characterization of OODs based on (Hsu et al. 2020) to deal with different kinds of reconstruction error distribution patterns. Concretely, we adopt a complexity score as a proxy measurement to quantify the "easiness" of OODs by off-the-shelf lossless image compression algorithm (Lin, Roy, and Li 2021; Serrà et al. 2019). As shown in Coefficient part of Figure 1, considering the essence of OOD detection is to find novel concepts at inference time, we further characterize semantic shift OOD to three kinds by the lower and upper complexity bound of ID training data removing extreme samples, i.e., the easiest and hardest top 5%. For example, the SVHN test images which have smaller complexities than lower complexity bound are categorized to easy OOD. Note that the test images whose complexities lie within the range of lower and upper complexity bound can be medium OOD or ID. Then, according to the type of inputs, we adjust and re-scale transformed reconstruction error with coefficient λ. We simply set λ to 0.5 for ID and keep the original reconstruction error for easy and hard OODs. Hence, the gap between easy, hard OODs and ID is enlarged. Besides, the finer characterization of OODs can serve as a principle to design equitable benchmark protocol. We notice that the experimental setup of dividing a multi-class dataset into ID and OOD adopted by (Ahmed and Courville 2020) and many OSR works is inadequate to evaluate OOD detector because they only consider medium OODs.

Inference

At inference time, an input image $\boldsymbol{x}$ and corresponding reconstruction $\hat{\boldsymbol{x}}$ are forward propagated through f_{fe}. For classification, the latent representation $\boldsymbol{z}$ is further propagated through f_{ch}. For OOD detection, we use two metrics under both scenarios to compute $Score_{\text{cla}}$ and $Score_{\text{rec}}$ in latent space, i.e., Mahalanobis distance and Euclidean distance. We call the two variants of our method READ-MD and READ-ED. Then, the $Score_{\text{rec}}$ is adjusted by coefficient λ based on image complexity. The final score is as follows:

$$Score = -Score_{\text{cla}} - \lambda * Score_{\text{rec}} \tag{10}$$

When the $Score$ is above a detection threshold τ, we assign the test input $\boldsymbol{x}$ as an ID sample. The above procedure is illustrated in the lower part of Figure 1.

Additionally, we adopt input perturbation strategy proposed in (Liang, Li, and Srikant 2017). They find that input perturbation brings larger gain on $Score$ for ID samples. We modify original strategy by perturbing over $Score_{\text{cla}} + Score_{\text{rec}}$. In detail, the perturbation of input $\boldsymbol{x}$ is given by:

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - \epsilon * \text{sign}(-\nabla_{\boldsymbol{x}}(Score_{\text{cla}}(\boldsymbol{x}) + Score_{\text{rec}}(\boldsymbol{x}, \hat{\boldsymbol{x}}))) \tag{11}$$

Then, $Score$ is recalculated with $\tilde{\boldsymbol{x}}$ and $\hat{\boldsymbol{x}}$ as described previously. Considering that test time OOD data is unavailable, the choice of hyperparameters depends on metric FPR@TPR95 of ID and synthesized OOD data from (Hendrycks, Mazeika, and Dietterich 2018), including uniform noise and 7 kinds corrupted ID samples, i.e., arithmetic mean, geometric mean, jigsaw, speckle noised, pixel, RGB ghosted, and inverted.

Experiments

In this section, we describe our experimental setup and demonstrate the effectiveness of our proposed method on various benchmark setups. Also, we conduct extensive ablation studies to explore different aspects of our algorithm. Code is publicly available at: https://github.com/lygjwy/READ.

Setup

In-distribution Datasets. CIFAR-10 (contains 10 classes) (Krizhevsky and Hinton 2009), and CIFAR-100 (contains 100 classes) (Krizhevsky and Hinton 2009) datasets are used as in-distribution data. We use the standard split, training set for training deep neural networks for image classification and reconstruction, and test set for evaluation.

Out-of-distribution Datasets. Considering the fine-grained characterization of OOD datasets, we use ten common benchmarks used in (Liang, Li, and Srikant 2017; Liu et al. 2020; Tack et al. 2020) for the comprehensiveness and fairness of evaluation as OOD test datasets: SVHN (Netzer et al. 2011), CIFAR-10, CIFAR-100, Textures (Cimpoi et al. 2014), Places365 (Zhou et al. 2017), TinyImageNet (crop) (Deng et al. 2009), TinyImageNet (resize) (Deng et al. 2009), LSUN (crop) (Yu et al. 2015), LSUN (resize) (Yu et al. 2015), and iSUN (Xu et al. 2015). In order to avoid overlapping with OOD validation data, we do not adopt uniform noise data. TinyImageNet (crop), TinyImageNet (resize), LSUN (crop), LSUN (resize), and

ID	OOD	FPR@95TPR ↓	AUROC ↑
		MSP / ODIN / Maha / Energy / READ-MD (ours)	
CIFAR-10	SVHN	48.3 / 33.2 / 15.3 / 35.4 / **12.0**	91.9 / 92.0 / 97.0 / 91.1 / **97.5**
	LSUN (c)	42.4 / 29.7 / 31.6 / **19.1** / 28.3	93.6 / 92.8 / 94.1 / **96.0** / 94.9
	Textures	59.5 / 49.5 / 18.0 / 52.5 / **10.3**	88.4 / 84.7 / 96.3 / 85.4 / **98.0**
	Places365	60.5 / 57.7 / 74.2 / **40.9** / 75.5	88.1 / 84.3 / 80.3 / **89.7** / 80.7
	CIFAR-100	62.9 / 60.7 / 71.8 / **50.5** / 76.5	**87.8** / 82.7 / 79.7 / 87.1 / 79.2
	TIN (c)	54.3 / 37.3 / 37.7 / 38.3 / **19.9**	90.5 / 91.6 / 92.9 / 91.5 / **96.5**
	LSUN (r)	52.0 / 26.5 / 34.1 / 27.9 / **9.4**	91.5 / 94.6 / 94.2 / 94.1 / **98.3**
	TIN (r)	60.8 / 39.1 / 34.1 / 46.5 / **12.3**	88.2 / 91.3 / 93.5 / 89.0 / **97.7**
	iSUN	56.4 / 32.4 / 33.5 / 33.9 / **12.5**	89.9 / 93.4 / 93.9 / 92.6 / **97.6**
	average	55.2 / 40.7 / 38.9 / 38.3 / **28.5**	90.0 / 89.7 / 91.3 / 90.7 / **93.4**
CIFAR-100	SVHN	85.0 / 82.1 / **58.0** / 92.2 / 67.9	70.3 / 69.1 / **85.3** / 73.6 / 81.8
	LSUN (c)	79.0 / 66.8 / 63.5 / 75.4 / **61.7**	77.6 / 81.2 / 82.0 / 83.1 / **83.1**
	Textures	83.1 / 78.8 / 36.9 / 78.0 / **35.6**	73.4 / 72.9 / 90.9 / 76.0 / **92.1**
	Places365	82.9 / 88.4 / 90.6 / **81.3** / 91.7	73.4 / 70.5 / 64.5 / **75.4** / 63.3
	CIFAR-10	**81.8** / 89.2 / 93.9 / 82.4 / 95.0	75.1 / 70.1 / 61.9 / **77.2** / 69.3
	TIN (c)	78.5 / 74.4 / 41.5 / 63.1 / **29.8**	76.5 / 80.0 / 91.0 / 81.2 / **93.6**
	LSUN (r)	82.5 / 73.9 / 22.7 / 62.0 / **10.9**	74.5 / 80.3 / 95.7 / 79.1 / **97.6**
	TIN (r)	82.3 / 71.6 / 25.3 / 63.5 / **14.7**	73.7 / 80.2 / 94.8 / 77.5 / **97.0**
	iSUN	83.1 / 70.6 / 26.2 / 62.3 / **15.5**	75.0 / 81.4 / 94.3 / 78.9 / **96.3**
	average	82.0 / 77.3 / 51.0 / 73.4 / **47.0**	74.4 / 76.2 / 84.5 / 78.0 / **84.9**

Table 1: Comparison with post-hoc methods. ↑ (↓) indicates larger (smaller) values are better. Bold numbers are superior.

iSUN are provided as a part of (Liang, Li, and Srikant 2017) code release.[1] Note that we preprocess cropped datasets with center clipping to remove the black border. We adopt officially original versions of the remaining datasets. For Places365, we use the same sampling as (Chen et al. 2021) for experimental results reproduction. The sampling list is publicly available at their code release.[2] All images are down-sampled to 32×32.

Networks and Training Details. We use WideResNet (Zagoruyko and Komodakis 2016), with depth 40, width 2 and dropout rate 0.3 as the classifier backbone. For READ-MD, we directly use the pre-trained classification model provided by (Liu et al. 2020) at their code release.[3] For READ-ED, we follow training details of (Hsu et al. 2020), the classifier is trained with batch size 128 for 200 epochs with weight decay 0.0005. The optimizer is SGD with momentum 0.9, and the learning rate starts with 0.1 and decreases by factor 0.1 at 50% and 75% of the training epochs. The weights in $h_i(\boldsymbol{x})$ of classifier are initialized with He-initialization (He et al. 2015) and not applied with weight decay. As for reconstruction model, we design an vanilla autoencoder with symmetrical structure, using ResNet18 (He et al. 2016) as encoder to deal with complex multi-class training data. The autoencoder is trained with batch size 128 for 2,000 epochs without weight decay. The optimizer is Adam with learning rate 0.001, betas 0.9 and 0.999. During training, we augment our training data with random flip and random cropping.

[1]https://github.com/facebookresearch/odin

[2]https://github.com/jfc43/informative-outlier-mining

[3]https://github.com/wetliu/energy_ood

Evaluation Metrics. We measure the following metrics: (1) the area under the receiver operating characteristic curve (AUROC); and (2) the false positive rate of OOD examples when true positive rate of ID data is at 95% (FPR@95TPR).

Results and Discussions

Main Results. The main results are reported in Table 1 and Table 2. For fair evaluation, we compare the proposed methods with competitive OOD detection algorithms which **do not rely on auxiliary OOD training data**. In Table 1, we show the performance of our method and other post-hoc methods based on discriminative models, including MSP (Hendrycks and Gimpel 2016), ODIN (Liang, Li, and Srikant 2017), Mahalanobis (Lee et al. 2018), and Energy (Liu et al. 2020). Over a total of 18 combinations of ID and OOD datasets, the proposed READ-MD algorithm outperforms the previous competing methods in 12 of them and gives second highest results on 2 of them.[4] Moreover, we show that using READ-MD reduces the average FPR@95TPR by **9.8%** compared to the second best Energy score when CIFAR-10 is ID, and **4.0%** compared to the second best Mahalanobis score when CIFAR-100 is ID. Without pre-training constraint, we present comparison results of the proposed READ-ED with three variants of G-OIDN (Hsu et al. 2020) in Table 2, i.e., G-ODIN-I, G-ODIN-C, and G-ODIN-E. Our method reduces the average FPR@95TPR by **1.5%** on ID CIFAR-10 compared to G-ODIN. The improvement is enlarged to **2.3%** on complex ID CIFAR-100. In particular, both of our methods decrease the FPR@95TPR metric for hard OODs by a large margin. It is worth not-

[4]This is based on the FPR@95TPR value; AUROC result is comparable.

ID	OOD	FPR@95TPR ↓	AUROC ↑
		G-ODIN-I / G-ODIN-C / G-ODIN-E / READ-ED (ours)	
CIFAR-10	SVHN	11.1 / 9.7 / **8.3** / 10.3	98.0 / 98.1 / **98.2** / 97.9
	LSUN (c)	6.1 / 11.0 / 3.1 / **2.8**	98.9 / 97.9 / 99.3 / **99.4**
	Textures	26.6 / 22.0 / 19.3 / **14.9**	94.9 / 96.0 / 96.7 / **97.4**
	Places365	42.0 / 34.1 / 25.8 / **25.7**	91.4 / 92.6 / 94.6 / **94.6**
	CIFAR-100	53.7 / 45.2 / 45.1 / **44.7**	88.3 / 89.9 / 90.7 / **90.8**
	TIN (c)	8.1 / 20.9 / 8.1 / **4.2**	98.5 / 96.2 / 98.5 / **99.1**
	LSUN (r)	3.0 / 13.4 / 2.7 / **1.3**	99.3 / 97.4 / 99.3 / **99.7**
	TIN (r)	6.2 / 24.0 / 8.6 / **4.5**	98.8 / 95.6 / 98.3 / **99.1**
	iSUN	2.8 / 16.1 / 2.7 / **1.5**	99.3 / 97.0 / 99.3 / **99.6**
	average	17.7 / 21.8 / 13.7 / **12.2**	96.4 / 95.6 / 97.2 / **97.5**
CIFAR-100	SVHN	65.6 / 78.2 / **36.6** / 63.9	85.2 / 83.6 / **94.0** / 89.5
	LSUN (c)	35.3 / 46.2 / **25.4** / 31.1	93.3 / 90.4 / **95.4** / 94.6
	Textures	80.0 / 40.7 / 21.7 / **17.9**	77.2 / 91.7 / 95.5 / **96.3**
	Places365	79.5 / **76.6** / 81.4 / 83.3	76.8 / **77.5** / 76.4 / 75.7
	CIFAR-10	**83.6** / 84.1 / 87.1 / 90.5	71.2 / **75.0** / 70.5 / 69.3
	TIN (c)	63.1 / 51.0 / 25.9 / **14.5**	87.1 / 90.1 / 95.3 / **97.5**
	LSUN (r)	75.6 / 56.7 / 22.9 / **6.5**	85.2 / 88.6 / 95.7 / **98.7**
	TIN (r)	73.5 / 51.0 / 20.6 / **7.9**	84.6 / 89.8 / 96.0 / **98.5**
	iSUN	78.6 / 57.0 / 24.7 / **10.5**	83.8 / 88.7 / 95.2 / **97.9**
	average	69.5 / 60.1 / 38.5 / **36.2**	82.7 / 86.1 / 90.4 / **90.9**

Table 2: Comparison with retraining methods. ↑ (↓) indicates larger (smaller) values are better. Bold numbers are superior.

ing that retraining classifier slightly deteriorate the classification performance, from 94.85% to 94.62% for CIFAR-10 and 75.83% to 75.08% for CIFAR-100.

Combination Study. To investigate how the performance of OOD detection changes when combining $Score_{\text{cla}}$ and $Score_{\text{rec}}$, we present detailed results for separated and aggregated OOD score in Table 3. Empirically, the combination brings lower FPR@95TPR and higher AUROC for most OOD datasets across our methods. The rationale is that the two scores represent the discrepancy of ID and OOD data from different perspectives and achieve an effect of complementation.

$-Score_{\text{cla}}$ / $-Score_{\text{rec}}$ / $-(Score_{\text{cla}} + Score_{\text{rec}})$		
Method	**FPR@95TPR ↓**	**AUROC ↑**
READ-MD	46.3 / 55.3 / **37.6**	90.2 / 75.4 / **90.8**
READ-ED	13.7 / 78.7 / **12.4**	97.2 / 59.1 / **97.5**

Table 3: OOD detection results for combination study. ↑ (↓) indicates larger (smaller) values are better. The results are averaged on nine OOD test datasets. Bold numbers are superior results.

Ablation Study. Table 4 validates the contributions of reconstruction error adjustment coefficient and input perturbation techniques. We report the average detection performance over 9 OOD datasets when CIFAR-10 is used as ID. After gradually applying techniques to our score function, one can note that reconstruction error adjustment decreases FPR@95TPR by **5.6%** for READ-MD. We do not present the ablation study result for READ-ED since the perturbation magnitude ϵ searched by ID and synthetic OOD data equals to 0. It is clear that the overconfidence problem for easy OOD of `AE` is alleviated. Therefore, the proposed adjustment coefficient is an indispensable part that strengthens our methods.

Method	Adj	Pert	FPR@95TPR ↓	AUROC ↑
READ-MD	-	-	37.6	90.8
	-	✓	29.9	92.7
	✓	-	33.3	92.3
	✓	✓	**28.5**	**93.4**

Table 4: OOD detection results for ablation study. ↑ (↓) indicates larger (smaller) values are better. Bold numbers are superior results. Adj and Pert mean adjustment and perturbation respectively.

Conclusion

In this work, we propose READ for out-of-distribution detection. The key idea is to unify distance to the closest class and reconstruction error in the latent space of classifier. We show that the combination of transformed reconstruction error exhibits superior detection performance. Against the overconfidence issue of autoencoder, we adjust the transformed reconstruction error with an image complexity based coefficient. As a result, the variants of READ, namely READ-MD and READ-ED, both achieve state-of-the-art performance in the corresponding scenario. Extensive ablations provide further understandings of our methods. We hope future work will pay more attention to mining and combining the inconsistencies of ID and OOD from different models.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (Grant No. 62192783, U1811462), the Collaborative Innovation Center of Novel Software Technology and Industrialization at Nanjing University.

References

Ahmed, F.; and Courville, A. 2020. Detecting semantic anomalies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3154–3162.

Chen, J.; Li, Y.; Wu, X.; Liang, Y.; and Jha, S. 2021. Atom: Robustifying out-of-distribution detection using outlier mining. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 430–445. Springer.

Choi, H.; Jang, E.; and Alemi, A. A. 2018. Waic, but why? generative ensembles for robust anomaly detection. *arXiv preprint arXiv:1810.01392*.

Cimpoi, M.; Maji, S.; Kokkinos, I.; Mohamed, S.; and Vedaldi, A. 2014. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3606–3613.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Denouden, T.; Salay, R.; Czarnecki, K.; Abdelzad, V.; Phan, B.; and Vernekar, S. 2018. Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance. *arXiv preprint arXiv:1812.02765*.

Gong, D.; Liu, L.; Le, V.; Saha, B.; Mansour, M. R.; Venkatesh, S.; and Hengel, A. v. d. 2019. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1705–1714.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hein, M.; Andriushchenko, M.; and Bitterwolf, J. 2019. Why ReLU Networks Yield High-Confidence Predictions Far Away From the Training Data and How to Mitigate the Problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hendrycks, D.; and Dietterich, T. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.

Hendrycks, D.; Mazeika, M.; and Dietterich, T. 2018. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*.

Hsu, Y.-C.; Shen, Y.; Jin, H.; and Kira, Z. 2020. 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*, 10951–10960.

Huang, R.; Geng, A.; and Li, Y. 2021. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 34.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *NeurIPS*.

Liang, S.; Li, Y.; and Srikant, R. 2017. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*.

Lin, Z.; Roy, S. D.; and Li, Y. 2021. Mood: Multi-level out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15313–15323.

Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33: 21464–21475.

Nalisnick, E.; Matsukawa, A.; Teh, Y. W.; Gorur, D.; and Lakshminarayanan, B. 2018. Do deep generative models know what they don't know? *arXiv preprint arXiv:1810.09136*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; and Snoek, J. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Oza, P.; and Patel, V. M. 2019. C2ae: Class conditioned auto-encoder for open-set recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2307–2316.

Ren, J.; Liu, P. J.; Fertig, E.; Snoek, J.; Poplin, R.; Depristo, M.; Dillon, J.; and Lakshminarayanan, B. 2019. Likelihood ratios for out-of-distribution detection. *Advances in Neural Information Processing Systems*, 32.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, 1278–1286. PMLR.

Sastry, C. S.; and Oore, S. 2020. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, 8491–8501. PMLR.

Sehwag, V.; Chiang, M.; and Mittal, P. 2021. SSD: A Unified Framework for Self-Supervised Outlier Detection. In *International Conference on Learning Representations*.

Serrà, J.; Álvarez, D.; Gómez, V.; Slizovskaia, O.; Núñez, J. F.; and Luque, J. 2019. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*.

Sun, Y.; Guo, C.; and Li, Y. 2021. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 34.

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

Van den Oord, A.; Kalchbrenner, N.; Espeholt, L.; Vinyals, O.; Graves, A.; et al. 2016. Conditional image generation with pixelcnn decoders. *Advances in neural information processing systems*, 29.

Xiao, Z.; Yan, Q.; and Amit, Y. 2020. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *Advances in neural information processing systems*, 33: 20685–20696.

Xu, P.; Ehinger, K. A.; Zhang, Y.; Finkelstein, A.; Kulkarni, S. R.; and Xiao, J. 2015. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*.

Yu, F.; Seff, A.; Zhang, Y.; Song, S.; Funkhouser, T.; and Xiao, J. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Zagoruyko, S.; and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhang, S.; Pan, C.; Song, L.; Wu, X.; Hu, Z.; Pei, K.; Tino, P.; and Yao, X. 2021. Label-Assisted Memory Autoencoder for Unsupervised Out-of-Distribution Detection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 795–810. Springer.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6): 1452–1464.