Improving OOD and OSR Detection in Traffic Sign Recognition using Wild Data

Ahmed Nassar
University of Rochester
anassar3@u.rochester.edu

Abstract

In conventional machine learning classification, our goal is to classify a datapoint into one of a predefined set of classes. In the real world, However, models often encounter data that does not belong to any of these classes, which can lead to arbitrary and unreliable predictions. That is the data is semantically different from our input distribution. This limitation arises because models lack the ability to recognize what they have not been trained to understand—they do not know what they do not know. For tasks with high stakes, such as autonomous driving, it is critical for models to detect when an input is out of distribution (OOD) and abstain from making an incorrect prediction. The task of identifying whether an input belongs to a different data distribution than the training set is referred to as Out-of-Distribution (OOD) Detection. A related but more challenging problem is Open Set Recognition (OSR), which focuses on detecting unknown labels within the same input distribution as the training data.

In this paper, we explore OOD detection and OSR in the context of traffic sign recognition. We create two datasets to evaluate our approach. The first dataset combines the German Traffic Sign Recognition Benchmark (GTSRB) with CIFAR-10 dataset, treating the CIFAR dataset as OOD data. The second dataset divides the GTSRB dataset into two halves: one half (30 classes) is used for training, while the remaining 13 classes are treated as the open set. We evaluate several detection methods, including max softmax probabilities, feature-based detection, and approaches incorporating wild (unlabeled) data during training. We evaluate the model using the following metrics: ROC, probabiltity density, and FPR@TPR95%.

Our results indicate that, for the task of traffic sign recognition, leveraging wild data during training significantly enhances the model's ability to detect OOD samples and recognize open-set instances. The code for this project is available at: GitHub Repository.

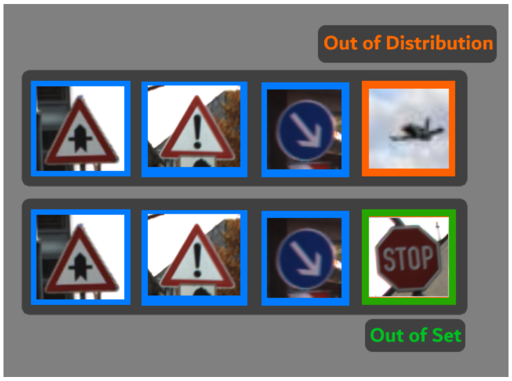


Figure 1. Demonstration of OOD and OOS

1. Introduction

Robust classifiers should not only produce accuracte predictions on their known contexts but they must also be able to predict when they see anomalies during inference. This is important as it improves the robustness of the model to changing input distributions. Such robustness is crucial in high-risk scenarios in the medical field such as classifying tumors or in autonomous driving such as classifing a novel object on the road. It is also helpful for the model to be able to recognize when it is seeing novel data as it could use that data later on for retraining during continual learning. In our case, we would like our model which is trained on the German Traffic sign recognition benchmark dataset. Conventional machine learning models are trained with the closed world assumption. meaning that that they expect the test data to be only drawn from the same distribution as the training data. several problems make the open world assumption. These are OOD detection Outlier detection. Anomaly detection, Novelity Detection, and Open Set Recognition. In our scenario, we are mostly concerned with Out of Distribution detection and Open Set Recognition.

The problem of traffic sign recognition is a subset of the more general problem of object recognition that is vital for autonomous driving. For the purpose of simplification, and in order to obtain smaller datasets and reduce training time, we analyze the smaller problem of traffic sign recognition hoping that our results should generalize to the problem of general object recognition. In the real world, a sensor on an autonomous vehicle is designed to recognize multiple classes of objects. using various machine vision techniques, it could recognize that this object is a traffic sign. if a similar looking object such as a Billboard or a fire hydrant is mistakenly classified as a traffic sign then we call that object as out of distribution. A more common scenario for our purposes would be mistaking a ”Stop” traffic sign for a ”turn” traffic sign due to the non-existance of a ”Stop” class in the training data. In this situation, the ”stop” class is in the same distribution but it is out of the training set. This problem can be divided into two primary categories: **Out-of-Distribution (OOD) Detection** and **Open Set Recognition (OSR)**.

Problem Formulation

Let the training dataset be defined as:

$$\mathcal{D}_{\text{train}} = \{(x_i, y_i) \mid x_i \in \mathcal{X}_{\text{train}}, y_i \in \mathcal{Y}_{\text{train}}\},$$

where:

- $\mathcal{X}_{\text{train}}$: Input space for the training data.
- $\mathcal{Y}_{\text{train}}$: Set of training labels, $\mathcal{Y}_{\text{train}} = \{1, 2, \ldots, C\}$, with C known classes.

During inference, the model encounters a new sample $x_{\text{test}} \in \mathcal{X}_{\text{test}}$, which may belong to:

1. **In-Distribution (ID):** x_{test} belongs to the same distribution as the training data, i.e., $x_{\text{test}} \sim \mathcal{P}(\mathcal{X}_{\text{train}})$, and its label $y_{\text{test}} \in \mathcal{Y}_{\text{train}}$.
2. **Out-of-Distribution (OOD):** x_{test} does not belong to the training data distribution, i.e., $x_{\text{test}} \not\sim \mathcal{P}(\mathcal{X}_{\text{train}})$. Its label $y_{\text{test}} \notin \mathcal{Y}_{\text{train}}$, and it is semantically different from the training classes (e.g., a fire hydrant mistaken as a traffic sign).
3. **Open Set:** $x_{\text{test}} \sim \mathcal{P}(\mathcal{X}_{\text{train}})$, but $y_{\text{test}} \notin \mathcal{Y}_{\text{train}}$. The data belongs to the same distribution as the training set but represents an unknown class (e.g., a ”Stop” sign when the model was only trained on ”Turn” signs).

Objective

The primary objective is to identify the optimal technique for OOD and OSR detection in the context of the traffic sign recognition problem. To evaluate, we rely on the FPR@TPR95% score as the primary evaluation metric, where the objective is to find the method that reduces this score the most.

1.1. Contributions

Our main contributions are increasing the applicability of traffic sig recognition systems beyond the closed-world assumption, as well as evaluating different detection techniques.

2. Background

Various works have attempted to tackle the problem of generalized OOD detection. Generally the methods can be classified as either inference-time detection or training-time regularization methods. inference-time techniques have the advantage of being easy to use without interfering with the training process which allows us to use pretrained models. However, they are generally less robust than training-time methods. Inference time detection can be done by assigning scores to the output of the model or to its internal features based on a distance function. Hendrycks et al, 2017 proposed a simple baseline method for OOD detection using the maximum softmax probabilites of the model. The assumption is that correctly classified examples tend to have greater maximum softmax probabilities than those incorrectly classified. However MSP scores are overconfident and tend to have a high False Positive Rate. Liang et al, 2018 proposed ODIN, a method that uses temperature scaling and input perturbations to enhance the seperability of OOD/ID data. However the score produced is no longer meaningful in terms of the confidence of the predictions. Liu et al, 2020 proposed an Energy-based OOD detection technique. The energy score is calculated as follows:

$$E(x) = -T \log \sum_i \exp^{f_i(x)/T} \tag{1}$$

Where T is a Temperature parameter. The energy score produced is less susceptible to the overconfidence issue of the baseline softmax probability method. Another type of inference-time detection are distance-based methods. Lee et al, 2018, use a score function based on the Mahalanobis distance of the Gaussian Distribution of the lower and upper-level features of the model. To calculate the Gaussian distribution we need to compute class mean $\hat{\mu_c}$ and covariance $\hat{\sum}$ of the training samples. Note that this does not require retraining as the mean and covariance can be computed from the feature embeddings of the model. Now the Mahalanobis score can be computed as:

$$M(x) = max_c - (f(x) - \hat{\mu_c} \hat{\sum}^{-1} (f(x) - \hat{\mu}_c) \tag{2}$$

The limitation of this approach is that it assumes input distributions will follow a gaussian distribution which is not always true.

Unlike inference time detection, Training time regularization techniques require changes during training time.

With training time regularization, instead of optimizing accuracy only on the **ID** data, we must also account for uncertainity from outside the **ID** data. formally, the objective of training in this case is:

$$argmin[R_{closed}(f) + \alpha * R_{open}(g)] \tag{3}$$

where R_{closed} is the classification error on the closed set or **ID** and R_{open} is the classification error on the open set or **OOD**. Liu et al, 2020, showed that this method improves seperability over the inference-based Energy-based detection method, however it requires auxillary data during training. Hendrycks et al, 2019 proposed Outlier Exposure as an approach to train models with auxilliary datasets. The approach is suitable for tasks in which auxiliary Data is abundant. Note that this Data does not need to be labeled so it is easy to procure it. we call such unlabeled Data as wild data.

Hyndrick's loss function can be computed as the sum of L_{ID} and a weight factor λ times L_{OOD}
where L_{ID} is computed as:

$$\mathcal{L}_{\text{ID}} = \frac{1}{N_{\text{ID}}} \sum_{i=1}^{N_{\text{ID}}} \text{CE}\left(f(x_i), y_i\right),$$

and L_{OOD} is computed as:

$$\mathcal{L}_{\text{OOD}} = \frac{1}{N_{\text{OOD}}} \sum_{j=1}^{N_{\text{OOD}}} \left(-\frac{1}{C} \sum_{k=1}^{C} z_j^k + \log \sum_{k=1}^{C} e^{z_j^k} \right),$$

where:

- z_j^k: The k-th logit of the model's output for the j-th OOD input.
- N_{OOD}: Number of OOD samples.
- C: Number of classes.
- $f(x_i)$: The model's output logits for the i-th in-distribution input.
- CE: Cross-entropy loss.
- N_{ID}: Number of in-distribution samples.

3. Method

3.1. Model Architecture and Training

While the focus of this project is not on developing a novel model architecture, we utilize a simple Convolutional Neural Network (CNN) model trained from scratch. This choice is motivated by the relatively small size of our datasets and the simplicity of the classification task, which does not require complex architectures. The model architecture consists of the following components:

- **Feature Extraction Layers**:
  - Two Conv2D layers, followed by a MaxPool2D layer and a Dropout layer.
  - Two additional Conv2D layers, followed by a MaxPool2D layer and a Dropout layer.
- **Classification Layers**:
  - A dense fully connected layer, followed by a Dropout layer.
  - A final dense layer for classification.

On the main in-distribution (ID) dataset, the German Traffic Sign Recognition Benchmark (GTSRB), this model achieves strong performance, with 96% accuracy on the test set after 10 epochs of training.

3.2. Datasets

We use two datasets to evaluate the robustness of our model:

1. **In-Distribution Dataset (ID)**:
   - **GTSRB**: Consists of 43 classes with 39,540 training images and 12,234 test images. This dataset is used to evaluate the model's classification performance and serves as the in-distribution data for Out-of-Distribution (OOD) and Open Set Recognition (OSR) experiments.
2. **Out-of-Distribution Dataset (OOD)**:
   - **CIFAR-10 Dataset**: Contains 50,000 images of various objects. This dataset is selected as OOD data due to its relatively small size and significant distributional difference from the GTSRB dataset.

3.3. Experimental Setup

We conduct two main evaluations: **OOD Detection** and **OSR Detection**.

3.3.1 Out-of-Distribution Detection (OOD)

To evaluate OOD detection, we combine the GTSRB dataset with the CIFAR-10 dataset. The model is trained on the GTSRB dataset, and its performance is tested on both GTSRB (ID) and the CIFAR dataset (OOD).

3.3.2 Open Set Recognition (OSR)

For OSR, we divide the GTSRB dataset into two subsets:

- A **closed set** consisting of 30 classes, which is used for training.

- An **open set** consisting of the remaining 13 classes, which is used for testing.

The model is trained on the closed set and evaluated on the entire 43-class dataset, representing the open set.

3.4. Detection Methods

We evaluate three detection methods in both OOD and OSR experiments:

3.4.1 Maximum Softmax Probabilities (MSP)

As a baseline, we use the maximum softmax probability (MSP) of the model's predictions to differentiate between ID and OOD samples. We visualize the results by plotting **probability density vs. MSP curves** for both experiments. Additionally, we generate **Receiver Operating Characteristic (ROC) curves** and compute the **FPR@TPR95% metric** using PyTorch's OOD library.

3.4.2 Feature-Based Detection (Mahalanobis Distance)

Using a feature-based approach, we compute the Mahalanobis distance between test samples and class-conditional Gaussian distributions derived from the training data. The PyTorch OOD library is used to compute metrics, and we generate:

- **Probability density vs. MSP curves.**

- **ROC curves.**

- **FPR@TPR95% scores.**

3.4.3 Training-Time Regularization (Outlier Exposure)

To improve OOD detection, we incorporate **Outlier Exposure** during training by modifying the loss function to penalize confident predictions on OOD samples. We evaluate this method by:

- Plotting **frequency vs. MSP curves** for both OOD and OSR experiments.

- Generating **ROC curves**.

- Computing the **FPR@TPR95% metric**.

This structured evaluation enables us to compare different detection methods and assess the robustness of our model across various scenarios.

4. Experiments

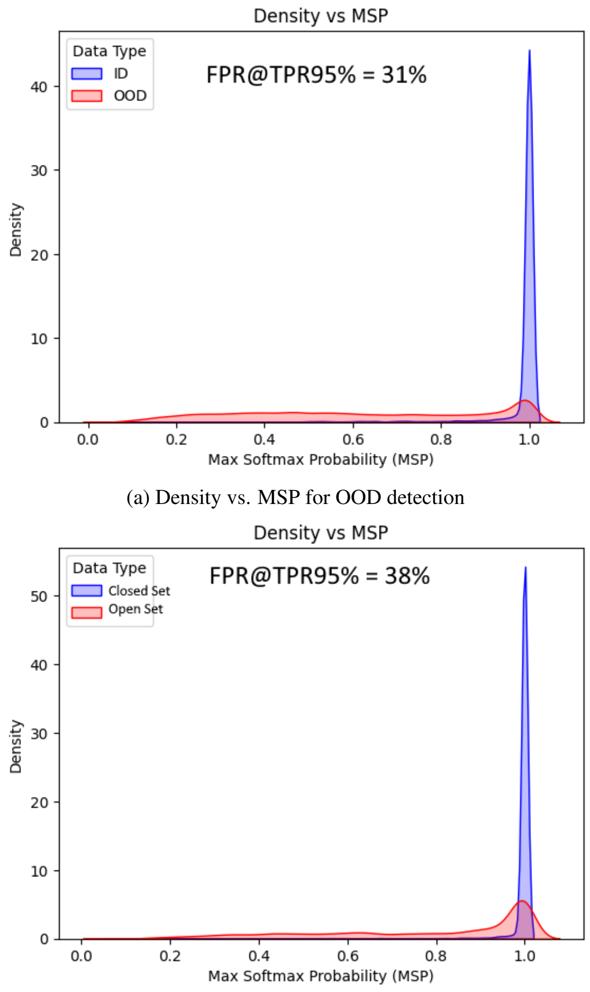
4.1. Maximum Softmax Probabilities (MSP) Baseline

For the first experiment, we evaluated the Maximum Softmax Probabilities (MSP) baseline for both Out-of-Distribution (OOD) detection and Open Set Recognition (OSR). We produced two sets of visualizations:

- Probability density vs. MSP curves.

- Receiver Operating Characteristic (ROC) curves.

These visualizations are shown in Figure 2 and Figure 3, respectively.

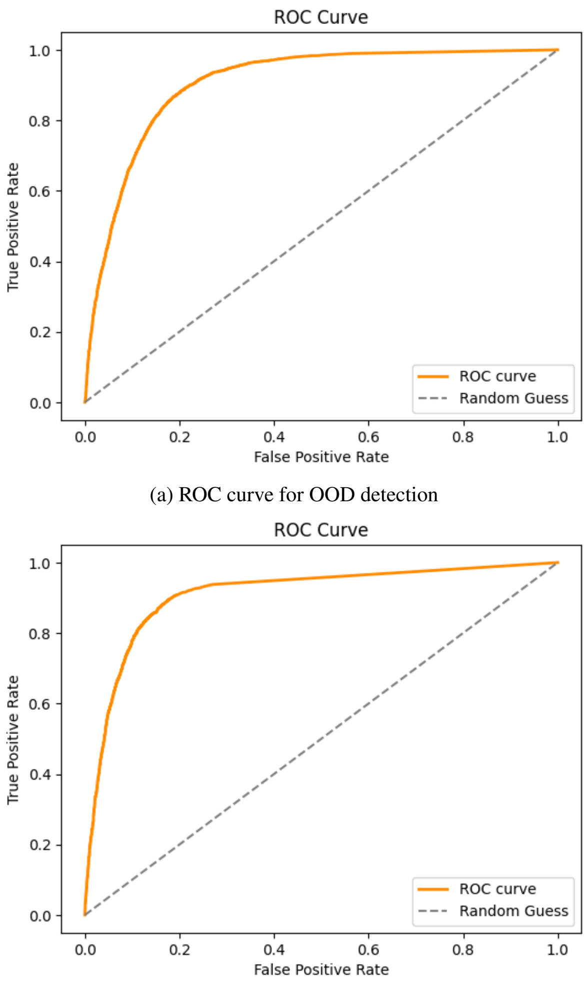


(a) Density vs. MSP for OOD detection

(b) Density vs. MSP for OSR detection

Figure 2. Density vs. MSP curves for MSP baseline

The results demonstrate that, as expected, OOD data is easier to distinguish than OSR data using the MSP baseline. However, both tasks exhibit high False Positive Rates



(a) ROC curve for OOD detection

(b) ROC curve for OSR detection

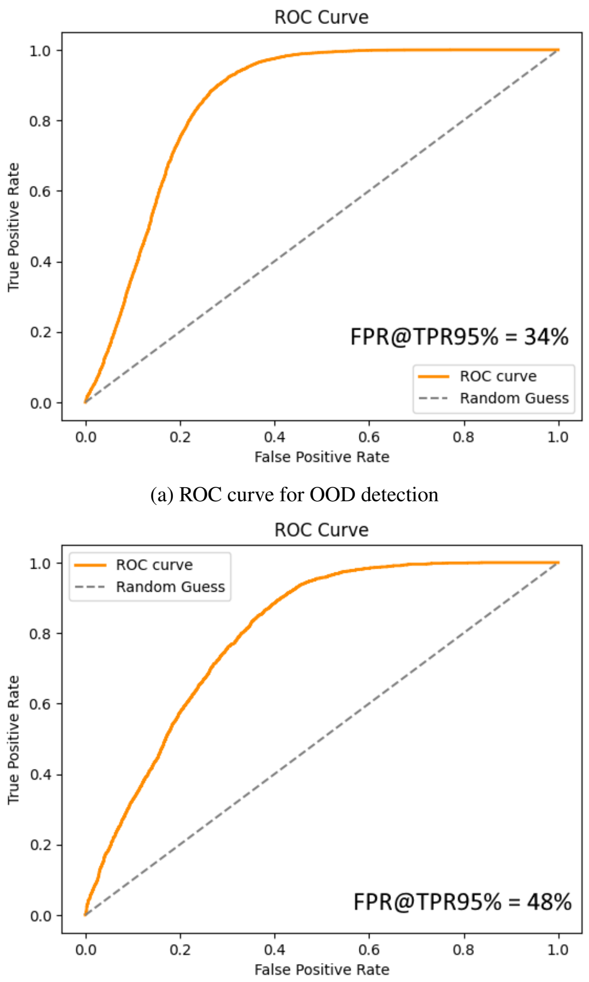
Figure 3. ROC curves for MSP baseline

(FPR) at a True Positive Rate (TPR) of 95%, highlighting the limitations of this approach.

4.2. Feature-Based Detection (Mahalanobis Distance)

Next, we performed feature-based detection using the Mahalanobis distance. Figure 4 shows the ROC curves for both OOD and OSR detection using this method.

The ROC curves indicate that the Mahalanobis distance performs worse than the MSP baseline in both OOD and OSR tasks, where the score of the FPR rate has especially been reduced down to 48%. This suggests that feature-based detection methods may not be well-suited for this specific scenario, yet further research need to be completed to understand why that is the case.



(a) ROC curve for OOD detection

(b) ROC curve for OSR detection

Figure 4. ROC curves for Mahalanobis distance

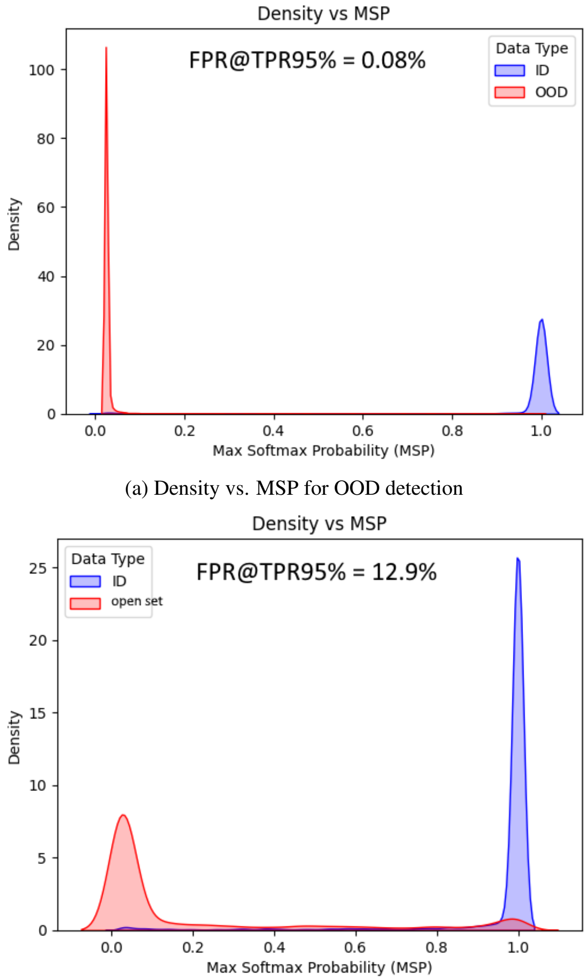
4.3. Training-Time Regularization (Outlier Exposure)

Finally, we evaluated the training-time regularization method using Outlier Exposure. This approach yielded excellent results for both OOD and OSR detection. Figures 5 and 6 show the respective density vs. MSP curves and ROC curves. on the OOD vs ID test, we acheived an $FPR@TPR95\%$ score of 0.08% , while the OSR score yeilded a respectable 12.9%

This method achieves significant improvements over the previous methods, reducing the FPR and achieving more robust performance in both OOD and OSR detection.

4.4. Conclusion

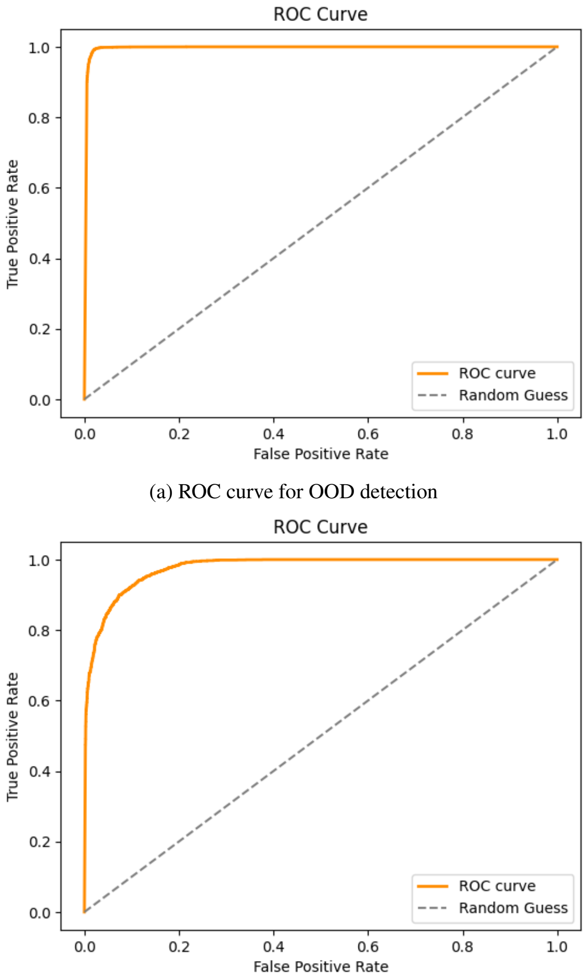
In this study, we evaluated three methods for OOD and OSR detection in the context of traffic sign recognition. Our experiments demonstrate that the training-time regulariza-



(a) Density vs. MSP for OOD detection

(b) Density vs. MSP for OSR detection

Figure 5. Density vs. MSP curves for Outlier Exposure



(a) ROC curve for OOD detection

(b) ROC curve for OSR detection

Figure 6. ROC curves for Outlier Exposure

tion approach, leveraging Outlier Exposure, provides the best results. For our specific scenario, we conclude that utilizing abundant unlabeled data from the real world during training is a highly effective strategy to improve robustness and detection capabilities.

References

[1] N. Drummond. The open world assumption. *eSI Workshop*, 2006.

[2] Xuefeng Du. Unknown-aware object detection: Learning what you don't know from videos in the wild.

[3] Dan Hendrycks. Deep anomaly detection with outlier exposure.

[4] Dan Hendrycks. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2017.

[5] Julian Katz-Samuels. Training ood detectors in their natural habitats.

[6] Kimin Lee. A simple unified framework for detecting out-of-distribution samples and adversarial attacks.

[7] Shiyu Liang. Enhancing the reliability of out-of-distribution image detection in neural networks. 2017.

[8] Weitang Liu. Energy-based out-of-distribution detection.

[9] Jingkang Yang. Generalized out-of-distribution detection: A survey.