A Comprehensive Survey on Out-of-Distribution Detection

1. Introduction to Out-of-Distribution Detection

1.1 Importance of Out-of-Distribution Detection

Out-of-distribution (OOD) detection has become increasingly crucial as machine learning (ML) models are being deployed in a wide range of real-world applications, including autonomous systems, medical diagnosis, and natural language processing. These applications often involve making critical decisions that can have significant consequences, making it essential to ensure the reliability and safety of the underlying ML models.

In the domain of autonomous systems, such as self-driving cars, OOD detection plays a vital role in ensuring the safety of passengers and pedestrians [1]. Without effective OOD detection capabilities, autonomous systems may make overconfident predictions on novel scenarios, leading to potentially catastrophic failures. For example, if a self-driving car's perception system fails to detect an unusual obstacle or traffic situation, the car may continue to drive at high speed, unable to react appropriately and avoid a collision. The risks associated with such failures can be severe, making OOD detection a critical component in the development of safe and reliable autonomous systems.

Similarly, in the medical domain, OOD detection is equally important for ensuring the accuracy and reliability of diagnostic tools [2]. Machine learning models trained on medical images or patient data may encounter novel cases that deviate from the training distribution, such as rare diseases, unusual symptoms, or atypical imaging artifacts. Failing to detect these OOD samples could lead to misdiagnosis or inappropriate treatment recommendations, with potentially life-threatening consequences for patients. By incorporating OOD detection capabilities, medical AI systems can flag anomalous cases and prompt further investigation or human review, improving the overall reliability and safety of these systems.

Likewise, in natural language processing (NLP) applications, such as chatbots, language translation, and text generation, OOD detection is crucial for maintaining the integrity and trustworthiness of the systems [3]. If an NLP model encounters input that is significantly different from its training data, it may generate nonsensical or inappropriate responses, potentially causing confusion or harm to users. OOD detection can help identify these atypical inputs and trigger appropriate actions, such as deferring to a human operator or providing a transparent indication of the model's uncertainty.

Beyond the specific application domains, the importance of OOD detection extends to the general deployment of machine learning models in the open world. As ML models are increasingly used to make decisions that impact people's lives, it is essential to ensure that these models can reliably distinguish between in-distribution (ID) data, which they are trained to handle, and OOD data, which may lead to unpredictable and potentially harmful outcomes [4]. Ignoring the potential for OOD samples can result

in overconfident and mistaken predictions, undermining the trustworthiness and reliability of the deployed systems.

To address these challenges, researchers have devoted significant effort to developing effective OOD detection methods [5]. These methods aim to equip ML models with the capability to identify and handle OOD samples, ensuring that the models can operate safely and reliably in diverse and unpredictable real-world environments. As the deployment of ML models continues to expand, the importance of OOD detection will only grow, making it a crucial area of research and development for building trustworthy and robust AI systems.

1.2 Defining Out-of-Distribution Detection

The out-of-distribution (OOD) detection problem refers to the task of identifying data samples that are significantly different from the training distribution, with the aim of ensuring the reliability and safety of machine learning models in real-world applications [6]. This challenge is closely related to, but distinct from, several other problems in machine learning, including anomaly detection, novelty detection, and open-set recognition.

Anomaly detection (AD) focuses on identifying data points that exhibit unusual patterns or deviations from the expected norm within the training data, without necessarily knowing the precise nature of the abnormal data [7]. In contrast, novelty detection (ND) aims to identify new or unknown data samples that are not present in the training distribution, by learning the characteristics of the training data and detecting significantly different samples [6].

Open-set recognition (OSR) extends the traditional classification task to handle unknown classes at test time, with the goal of correctly classifying samples from known classes while also identifying samples that belong to unknown or novel classes [8]. This is particularly relevant in scenarios where the set of classes encountered during deployment may be larger or different from the set used during training.

While these related problems share some common underlying principles, they have distinct characteristics and focus on different aspects of out-of-distribution detection. OOD detection, in its broader sense, encompasses all these tasks and aims to develop robust and reliable methods for identifying samples that deviate from the training distribution, regardless of the specific nature of the deviation [9].

One key distinction is the availability of labeled data for the out-of-distribution or novel samples. Anomaly detection and novelty detection typically rely on learning from only the normal or in-distribution samples, while open-set recognition and some OOD detection approaches may have access to a limited number of labeled examples of the out-of-distribution or unknown classes during training [8]. Another difference lies in the focus on different types of distribution shifts, with anomaly detection and novelty detection often addressing changes in the underlying data characteristics, and open-set recognition and OOD detection also considering shifts in the semantic content of the data.

Despite these differences, these related problems can be unified under a broader framework of OOD detection, where the goal is to identify samples that are significantly different from the training distribution, regardless of the specific nature of the distribution shift [9]. This unified perspective can help in developing more comprehensive and generalizable solutions that can handle a diverse range of out-of-distribution scenarios.

1.3 Challenges in OOD Detection

Designing effective out-of-distribution (OOD) detection methods faces several key challenges. A primary challenge is the lack of OOD data during the training phase. Unlike in-distribution (ID) data, which is typically readily available, OOD data can be diverse and difficult to obtain [6]. This poses a significant obstacle, as models trained exclusively on ID data may struggle to generalize and accurately detect OOD samples during deployment.

Another crucial challenge is the need to handle diverse types of OOD distributions. OOD samples can differ from ID data in various ways, such as semantic shifts (e.g., a bird image in a dog classification task) or covariate shifts (e.g., a blurry image in a sharp image classification task) [10]. Successful OOD detectors must be able to identify these diverse distribution shifts and distinguish them from the ID data. This requires the development of robust and flexible models that can adapt to a wide range of OOD scenarios.

Furthermore, there is often a trade-off between OOD detection and in-distribution classification performance. Methods that focus solely on improving OOD detection may compromise the classification accuracy on ID data, making them less suitable for real-world deployment [11]. Conversely, models optimized for ID classification may not be able to effectively detect OOD samples. Striking the right balance between these two objectives is crucial for the practical application of OOD detection techniques.

In addition to the conceptual challenges, there are also practical considerations, such as the computational efficiency requirements for real-world deployment. OOD detection methods need to be lightweight and perform inference quickly, especially in time-sensitive applications like autonomous driving [1]. Computationally expensive approaches or those that require extensive pre-processing or fine-tuning may not be suitable for deployment in resource-constrained environments.

Another challenge arises from the fact that OOD detection is inherently a domain-specific problem, as the definition of "out-of-distribution" can vary depending on the application and the underlying data [10]. What may be considered OOD in one context may be in-distribution in another. Developing generalizable OOD detection methods that can adapt to different domains and tasks is a significant challenge.

Furthermore, the evaluation of OOD detection methods is itself a complex task. Existing benchmarks and evaluation protocols often have

limitations, such as the use of dataset biases or the lack of realistic distribution shifts [11]. Designing comprehensive and representative evaluation frameworks is crucial for assessing the true performance and generalization capabilities of OOD detection methods.

In summary, the main challenges in designing effective OOD detection methods include the lack of OOD data during training, the need to handle diverse types of OOD distributions, the trade-off between OOD detection and in-distribution classification performance, the computational efficiency requirements for real-world deployment, the domain-specific nature of OOD detection, and the limitations of existing evaluation protocols. Overcoming these challenges requires innovative approaches, careful model design, and the development of more comprehensive and realistic benchmarking frameworks.

1.4 Evaluation of OOD Detection

Evaluating the performance of out-of-distribution (OOD) detection methods is crucial for understanding their effectiveness and practical applicability. Existing research has predominantly relied on a set of commonly used benchmark datasets and evaluation metrics, although these approaches have several limitations and biases.

One of the key challenges in evaluating OOD detection methods is the lack of diversity and potential biases in the benchmark datasets. The widely used CIFAR-10, CIFAR-100, and ImageNet datasets, while providing a standard testbed, may not capture the complexity and real-world distribution shifts that OOD detectors are expected to handle [12; 5]. For instance, the low-resolution images and limited object categories in CIFAR-10 and CIFAR-100 may not adequately represent the diverse OOD scenarios encountered in real-world applications [13].

Similarly, the commonly used evaluation metrics, such as the area under the receiver operating characteristic (AUROC) curve and the false positive rate at 95% true positive rate (FPR95), have been criticized for their focus on overall performance, without providing insights into the specific types of distribution shifts or characteristics of OOD samples that the detectors struggle with [14]. This can lead to a skewed perception of the real-world performance of OOD detectors [12].

To address these limitations, researchers have proposed more comprehensive and realistic evaluation setups for OOD detection. For example, the Towards Realistic Out-of-Distribution Detection study introduced new OOD test datasets, such as CIFAR-10-R, CIFAR-100-R, and ImageNet-30-R, which exhibit more diverse and realistic distribution shifts [13]. These datasets aim to better simulate the challenges faced by OOD detectors in real-world scenarios, where the distribution of the data may change over time or across different deployment environments.

Additionally, the Unsupervised Evaluation of Out-of-distribution Detection study proposed a novel approach for evaluating OOD detection methods in the absence of ground truth labels for OOD samples, introducing the Gscore, an unsupervised indicator of OOD detection

performance [15]. This can be particularly useful when ground truth OOD labels are not available.

Furthermore, the OpenOOD framework provides a more comprehensive benchmark for evaluating OOD detection methods, encompassing not only traditional OOD detection but also related tasks such as anomaly detection, novelty detection, and open-set recognition [5]. This unified evaluation framework allows for a more systematic comparison of different OOD detection approaches and their applicability across a wide range of real-world scenarios.

By addressing the limitations of current evaluation approaches and introducing more diverse and challenging benchmark datasets, the research community is working towards developing a better understanding of the performance of OOD detectors in real-world applications. This will ultimately lead to the design of more reliable and robust OOD detection systems that can be deployed in safety-critical domains with confidence.

1.5 Applications and Practical Considerations

The deployment of out-of-distribution (OOD) detection methods in real-world applications faces several practical challenges that must be addressed to ensure their effectiveness and reliability. One key challenge is the need for diverse training data that can capture the wide range of scenarios the model will encounter during deployment [6]. In many applications, the available training data may not fully represent the complexity and diversity of the real-world, leading to poor OOD detection performance as the model struggles to identify samples that deviate from the limited training distribution.

To address this challenge, researchers have emphasized the importance of developing OOD detectors that can adapt to dynamic and evolving distributions [16]. This requires the OOD detection models to be robust to both covariate shifts, which refer to changes in the input feature distribution, and semantic shifts, which involve changes in the output label distribution [10]. Developing OOD detectors that can handle both types of distribution shifts is crucial for ensuring their reliable performance in real-world deployments [17].

Another practical consideration in deploying OOD detection methods is computational efficiency. Many state-of-the-art OOD detection techniques, such as those based on deep learning, can be computationally intensive, making them challenging to deploy in resource-constrained environments [18]. This is particularly relevant for applications in the time series domain, where the detection needs to be performed in real-time, such as in autonomous driving [19] and healthcare monitoring [20].

To address these computational challenges, researchers have explored techniques like test-time adaptation, which can leverage unlabeled data during inference to enhance OOD detection performance [21]. Additionally, some studies have investigated ways to optimize the detection models for efficient deployment, such as by using non-parametric approaches [22] or by leveraging the properties of pre-trained language models [23].

The real-world applications of OOD detection span various domains, including computer vision, natural language processing, and time series analysis. In computer vision, OOD detection is crucial for the reliability of object detection systems in autonomous vehicles [24] and for ensuring the safety of medical imaging applications [25]. In natural language processing, OOD detection can help identify anomalous inputs that could lead to unreliable predictions in tasks such as text classification [26] and dialogue systems.

In the time series domain, OOD detection is essential for monitoring sensor data in industrial applications [27] and for ensuring the reliability of predictive models in healthcare [20]. By identifying anomalous samples that deviate from the training distribution, OOD detection can help prevent catastrophic failures and enhance the overall safety and reliability of these systems.

1.6 Emerging Trends and Future Directions

The field of out-of-distribution (OOD) detection has seen remarkable progress in recent years, with researchers proposing a wide range of techniques to tackle this crucial challenge. However, as the complexity of machine learning models and their deployment environments continue to evolve, several open challenges and promising future research directions have emerged.

One of the key challenges in OOD detection is dealing with distribution shifts, which can occur due to various factors, such as changes in the environment, sensor degradation, or the introduction of new data sources. Existing OOD detection methods often rely on the assumption that the test distribution is similar to the training distribution, but this assumption may not hold in real-world scenarios [28]. To address this issue, future research should focus on developing OOD detectors that are more robust and adaptive to dynamic and evolving distributions, potentially by leveraging techniques from the domain adaptation and domain generalization literature [17].

Another promising avenue for future research is the leveraging of unlabeled data to improve OOD detection performance. While the majority of existing OOD detection methods rely on labeled in-distribution data, the availability of such data may be limited in practice. Researchers have recently proposed methods that utilize unlabeled data, such as separating candidate outliers from unlabeled data [29] and learning with a mixture of prototypes to capture the diversity of in-distribution data [30]. These approaches demonstrate the potential of unlabeled data to enhance OOD detection, and further investigation in this direction could lead to more data-efficient and generalizable solutions.

In addition to distribution shifts and the use of unlabeled data, there is also a growing interest in integrating OOD detection with other machine learning tasks, such as open-set recognition, anomaly detection, and few-shot OOD detection [9]. By taking a more holistic approach to OOD detection, researchers can leverage the synergies between these related problems and develop more comprehensive solutions that can handle a broader range of scenarios.

Finally, the development of interpretable and robust OOD detectors is another important direction for future research. While current OOD detection methods have demonstrated promising results, many of them remain opaque in their decision-making process, making it difficult to understand and trust their predictions. Researchers should explore techniques that can provide insights into the factors contributing to OOD detection, such as the use of saliency maps or concept-based explanations [31]. Additionally, the robustness of OOD detectors to adversarial attacks and other forms of distributional shifts remains a critical concern, and future work should aim to develop methods that are more resilient to such challenges [32].

2. Taxonomy and Formulation of OOD Detection

2.1 Taxonomy of Out-of-Distribution Detection Problems

The problem of detecting out-of-distribution (OOD) data has been studied under various names in the machine learning community, including anomaly detection, novelty detection, open-set recognition, and OOD detection. While these problems share some common aspects, they differ in their specific problem settings and evaluation metrics.

Anomaly detection [33] focuses on identifying data points that deviate significantly from the majority of the training data, which is assumed to represent the "normal" or in-distribution (ID) samples. The goal is to detect anomalous instances that may be indicative of errors, failures, or other rare events. Anomaly detection methods often rely on statistical models or proximity-based approaches to identify outliers in the data.

In contrast, novelty detection [34] aims to identify new or previously unseen data that is distinct from the training distribution. Unlike anomaly detection, novelty detection methods do not assume that the training data is homogeneous or representative of the entire ID distribution. Instead, they focus on identifying samples that are dissimilar to the known classes in the training set.

Open-set recognition [35] extends the traditional closed-world classification problem by allowing for the possibility of unknown classes at test time. In this setting, the model must not only classify samples into the known classes but also detect and reject instances that belong to unknown classes. This is particularly relevant in real-world scenarios where the set of classes encountered during deployment may not be fully known during the training phase.

Out-of-distribution (OOD) detection [3] is a broader problem that encompasses both anomaly detection and novelty detection. The goal of OOD detection is to identify test samples that are significantly different from the training distribution, regardless of whether they belong to known or unknown classes. This problem is crucial for the safe deployment of machine learning models in open-world applications, where the model may encounter a wide range of inputs that were not present in the training data.

While these problems share some similarities, they differ in their specific problem formulations and evaluation metrics. Anomaly detection is often evaluated using metrics such as the area under the receiver operating characteristic (AUROC) curve and the false positive rate at a fixed true positive rate (FPR95). Novelty detection and open-set recognition typically use similar metrics, but may also incorporate measures of the model's ability to correctly classify known classes.

For OOD detection, the evaluation metrics often include AUROC, FPR95, and the area under the precision-recall curve (AUPR). Additionally, some studies have proposed more comprehensive evaluation protocols, such as the OpenOOD benchmark [5], which considers different types of OOD data and evaluates methods under various settings.

Despite the differences in problem formulations and evaluation metrics, these research areas share a common goal of enhancing the reliability and robustness of machine learning models in the face of unfamiliar or unexpected data. As such, cross-pollination of ideas and techniques between these fields can lead to advancements in the broader problem of detecting and handling OOD samples.

2.2 Anomaly Detection

Anomaly detection is a long-standing research area in machine learning that aims to identify data points that are significantly different from the normal samples in the training data. The goal is to detect observations that do not conform to an expected pattern or behavior, often referred to as "anomalies", "outliers", or "novelties". Anomaly detection has a wide range of applications, including fraud detection, intrusion detection, fault diagnosis, and medical diagnosis [6].

While anomaly detection is related to the broader problem of out-of-distribution (OOD) detection, it differs in its specific problem formulation and evaluation metrics. In the context of OOD detection, anomaly detection can be viewed as a special case where the objective is to identify samples that are significantly different from the in-distribution data used for training the model. This is particularly important in safety-critical applications, where the deployment of machine learning models in the open world needs to handle unexpected or unseen inputs that could lead to unreliable or unsafe predictions.

Traditionally, anomaly detection methods have focused on learning a model of the normal data distribution and then using this model to identify samples that deviate from the learned normality. These methods can be broadly categorized into statistical techniques, proximity-based approaches, and reconstruction-based methods [36]. Statistical techniques, such as Gaussian mixture models and one-class support vector machines, aim to estimate the likelihood or density of the normal data and flag low-likelihood samples as anomalies. Proximity-based approaches, such as k-nearest neighbors and isolation forests, identify anomalies as data points that are far away from the majority of the training samples. Reconstruction-based methods, such as autoencoders and generative adversarial networks, learn a compressed representation of the normal data and use the reconstruction error as an anomaly score.

Recent advancements in deep learning have led to the development of more powerful anomaly detection techniques that can handle complex, high-dimensional data, such as images and time series [36]. These deep anomaly detection methods often leverage the feature representations learned by deep neural networks to identify anomalies. For example, some methods use the logits or activations of a pre-trained classifier as the anomaly score, under the assumption that in-distribution samples will have higher classification confidence compared to out-of-distribution samples [6].

However, a key challenge in anomaly detection is the lack of labeled anomaly data during training, as anomalies are inherently rare and difficult to obtain. To address this, some deep anomaly detection methods have explored self-supervised learning techniques, where the model is trained to learn useful representations from the normal data without the need for explicit anomaly labels [37]. These methods often employ generative adversarial networks or contrastive learning approaches to capture the underlying structure of the normal data and identify outliers based on their deviation from this learned representation.

Furthermore, recent studies have also highlighted the importance of considering the distribution shift between the training and test data, as anomalies can arise not only from semantic differences but also from covariate shifts [38]. This has led to the development of more robust anomaly detection methods that can handle diverse types of distribution shifts, including label noise and dataset bias.

In summary, anomaly detection is a crucial component of OOD detection, as it aims to identify data points that are significantly different from the normal samples in the training data. The field has seen significant advancements in recent years, particularly with the adoption of deep learning techniques and the consideration of distribution shifts. However, challenges remain in handling the lack of labeled anomaly data and developing anomaly detectors that are both effective and robust to various types of distribution shifts.

2.3 Novelty Detection

Novelty detection is a fundamental problem in machine learning, with applications ranging from anomaly detection in industrial processes to identifying new biological species. In the context of out-of-distribution (OOD) detection, novelty detection refers to the task of identifying data samples that are not present in the training distribution. Unlike anomaly detection, which aims to identify samples that are significantly different from the normal data, novelty detection focuses on identifying new or unknown samples that are qualitatively distinct from the in-distribution data.

One of the key challenges in novelty detection is the lack of labeled data for the novel or unknown classes, as by definition, these samples are not present in the training data. To address this issue, researchers have explored a range of unsupervised and semi-supervised approaches for novelty detection. A common approach is to learn a generative model of the in-distribution data, such as a variational autoencoder (VAE) or a

generative adversarial network (GAN), and then use the model's reconstruction error or likelihood as a measure of novelty [6]. The idea is that samples from the in-distribution should be well-reconstructed by the generative model, while novel samples will have high reconstruction errors. This approach has been extended to incorporate additional signals, such as the latent representations of the generative model, to improve the novelty detection performance [39].

Another popular approach to novelty detection is to learn a one-class classifier, where the goal is to learn a decision boundary that separates the in-distribution data from the novel or unknown samples [9]. This can be achieved through techniques such as one-class support vector machines (OC-SVMs) or deep one-class classification methods. These approaches aim to learn a compact representation of the in-distribution data and use this representation to identify novel samples that lie outside the learned decision boundary.

More recently, there has been a growing interest in leveraging self-supervised learning techniques for novelty detection. By learning representations that capture the underlying structure of the in-distribution data, self-supervised methods can be effective in distinguishing novel samples from the known data distribution [40]. For example, contrastive learning approaches have been shown to be effective in learning representations that are invariant to certain transformations, which can be useful for identifying samples that do not adhere to these transformations.

Despite the progress made in novelty detection, there are still several open challenges and research directions. One key challenge is the need for more comprehensive and realistic evaluation protocols for novelty detection, as many existing benchmarks may not capture the full range of real-world scenarios [13]. Additionally, there is a need for more interpretable and explainable novelty detection methods, which can provide insights into the decision-making process and the characteristics of the novel samples [31]. Another important research direction is the integration of novelty detection with other machine learning tasks, such as few-shot learning or open-set recognition [23], which may lead to more robust and versatile novelty detection systems that can handle a wider range of real-world challenges.

2.4 Open-Set Recognition

Open-set recognition is a variant of the traditional classification task that aims to handle unknown classes at test time. Unlike the closed-set classification scenario, where the model is trained and evaluated on a predefined set of classes, open-set recognition introduces the challenge of detecting and classifying instances that belong to classes not seen during training [41].

This challenge is particularly relevant in real-world applications, where it is often impractical or infeasible to collect training data for every possible class that the model may encounter. This could be due to the sheer number of potential classes, the difficulty in obtaining representative samples for all classes, or the fact that new classes may

emerge over time [41]. In such scenarios, a model trained on a fixed set of classes needs to be able to not only classify instances belonging to those classes but also detect and handle instances from unknown or unseen classes.

The key distinction between open-set recognition and the more extensively studied problem of out-of-distribution (OOD) detection is that open-set recognition aims to classify instances into either known or unknown classes, whereas OOD detection focuses solely on identifying whether an instance is from the training distribution or not [42]. This added complexity of open-set recognition presents unique challenges in both the model design and the evaluation process.

From a model design perspective, open-set recognition requires the model to learn a decision boundary that can effectively separate known classes from unknown classes, rather than just a boundary between different known classes. This often involves incorporating mechanisms to model the uncertainty or confidence of the model's predictions, as well as techniques to leverage the available training data more efficiently [43]. Additionally, open-set recognition models need to be robust to the potentially large and diverse distribution of unknown classes, which can be significantly different from the known classes.

In terms of evaluation, open-set recognition introduces new metrics and benchmarks that go beyond the traditional classification accuracy. These include measures such as open-set recognition accuracy, which considers both the model's ability to correctly classify known instances and its ability to detect unknown instances [41]. Designing appropriate evaluation protocols and benchmark datasets is crucial for assessing the performance of open-set recognition models in a reliable and meaningful way.

Recent advances in open-set recognition have leveraged various techniques, such as generative models, meta-learning, and adversarial training, to improve the model's ability to handle unknown classes [44]. However, significant challenges remain, particularly in scaling open-set recognition to large-scale, real-world scenarios with diverse and dynamic class distributions [45].

Overall, open-set recognition is an important and active area of research that aims to bridge the gap between the controlled settings of traditional classification and the open-world challenges faced by real-world AI systems. The continued development of robust and reliable open-set recognition models will be crucial for the safe and trustworthy deployment of machine learning in a wide range of applications.

2.5 Evaluation Metrics

Evaluating the performance of out-of-distribution (OOD) detection methods is a crucial aspect of understanding their capabilities and limitations. To assess the effectiveness of OOD detectors, the research community has established several standard evaluation metrics that capture different aspects of the detector's performance.

One of the most widely used metrics is the Area Under the Receiver Operating Characteristic (AUROC) curve, which measures the trade-off between the true positive rate (TPR) and the false positive rate (FPR) across different detection thresholds [46; 47; 48]. A higher AUROC indicates better OOD detection capabilities.

Another commonly used metric is the False Positive Rate at 95% True Positive Rate (FPR95) [12; 47; 48]. This metric focuses on the detector's ability to maintain a low false positive rate while maintaining a high true positive rate, which is particularly relevant in safety-critical applications.

In addition to AUROC and FPR95, researchers have employed other evaluation metrics, such as the Area Under the Precision-Recall Curve (AUPR) [46], which captures the trade-off between precision and recall. The Negative Log-Likelihood (NLL) [49] has also been used as a metric for OOD detection methods based on likelihood estimation.

It is important to note that the choice of evaluation metrics can significantly impact the assessment of OOD detection methods, as different metrics capture different aspects of performance [12]. Researchers have recognized the need for comprehensive and realistic evaluation protocols for OOD detection, leading to the development of new datasets and evaluation frameworks [47; 50; 47] to better assess the performance of OOD detectors in more challenging and diverse settings.

3. Methodologies for OOD Detection

3.1 Classification-based Approaches

Out-of-distribution (OOD) detection is a crucial task for ensuring the reliability and safety of machine learning systems, especially in safety-critical applications. Among the various approaches proposed for OOD detection, classification-based methods have been extensively studied and have shown promising results.

One of the earliest and most widely used classification-based OOD detection methods is the maximum softmax probability (MSP) approach proposed by Hendrycks and Gimpel [3]. The key idea behind MSP is to leverage the softmax output of a pre-trained classification model to detect OOD samples. The underlying assumption is that in-distribution (ID) samples will have a higher maximum softmax probability compared to OOD samples, as the model is more confident in its prediction for the former. By setting a suitable threshold on the maximum softmax probability, the model can distinguish between ID and OOD samples.

While the simplicity and ease of implementation of the MSP approach make it an attractive option, it has been shown to have limitations in certain scenarios. For example, [51] demonstrated that MSP can perform poorly when the OOD samples are close to the decision boundary of the classifier, leading to high false positive rates.

To address the limitations of the MSP approach, several extensions and modifications have been proposed. One such method is the ODIN (Outlier

Detection In Neural Networks) approach, introduced by Liang et al. [52]. ODIN introduces two key modifications to the MSP approach: temperature scaling and input preprocessing. Temperature scaling adjusts the softmax temperature to increase the separation between the maximum softmax probability of ID and OOD samples, while input preprocessing applies small perturbations to the input to amplify the differences in softmax outputs between ID and OOD samples. These modifications have been shown to improve the performance of OOD detection compared to the vanilla MSP approach.

Another classification-based approach is the Mahalanobis distance-based method proposed by Lee et al. [51]. This method leverages the Mahalanobis distance between the input sample and the class-conditional Gaussian distributions learned from the ID data to detect OOD samples. The Mahalanobis distance-based method has been shown to outperform the MSP and ODIN approaches, particularly in scenarios where the OOD samples are close to the decision boundary.

More recently, the Deep Anomaly Detection (DAD) framework, proposed by Ruff et al. [53], has gained attention in the OOD detection literature. DAD learns a deep neural network to map the input data to a latent space, where a one-class classifier is trained to detect anomalies (i.e., OOD samples). By leveraging the representational power of deep neural networks and the flexibility of one-class classifiers, DAD has been shown to outperform traditional OOD detection methods in various benchmarks.

Furthermore, the emergence of large language models (LLMs) [48] has also sparked interest in exploring their capabilities for OOD detection in the text domain. These powerful pre-trained models have shown promising results in detecting OOD samples, often outperforming specialized OOD detection methods.

In summary, classification-based approaches for OOD detection have evolved from the simple MSP method to more sophisticated techniques, such as ODIN, Mahalanobis distance-based methods, and deep anomaly detection frameworks. These methods leverage the classification model's output, either directly or through various modifications, to distinguish between ID and OOD samples. While these approaches have shown promising results, they often rely on the assumption that the OOD samples are significantly different from the ID samples, which may not always hold in real-world scenarios. Ongoing research aims to address these limitations and further improve the robustness and generalization of classification-based OOD detectors.

3.2 Density-based Approaches

Density-based approaches for out-of-distribution (OOD) detection aim to estimate the likelihood or density of input samples and use this information to identify OOD data. These methods generally assume that in-distribution (ID) samples are drawn from a well-defined underlying distribution, while OOD samples are more likely to come from a different, less-defined distribution.

One class of density-based OOD detection methods are the likelihood-based approaches, which directly estimate the likelihood or probability density of the input samples. These methods often utilize generative models, such as variational autoencoders (VAEs) [6] or normalizing flows [6], to model the distribution of ID samples. The likelihood or probability density of a test sample is then used as an OOD indicator, with low-likelihood samples being classified as OOD.

For example, ODIN [6] is a likelihood-based method that leverages the output logits of a pre-trained classification model to estimate the likelihood of input samples. It applies temperature scaling and input perturbations to amplify the differences in logits between ID and OOD samples, thereby improving the OOD detection performance.

Another density-based approach is the use of energy-based models (EBMs) [6], which learn an energy function that assigns low energy to ID samples and high energy to OOD samples. The energy of a test sample is then used as the OOD indicator, with high-energy samples being classified as OOD.

In addition to likelihood-based methods, flow-based approaches have also been explored for OOD detection. These methods use normalizing flow models, which are generative models that learn a bijective mapping between the input space and a latent space with a simple, tractable probability distribution. By estimating the log-likelihood of the input sample under the flow model, flow-based approaches can identify OOD samples as those with low likelihood.

For instance, Gaussian Mixture FLOW (GM-FLOW) [6] is a flow-based OOD detection method that models the in-distribution using a Gaussian mixture model in the latent space of a normalizing flow. The log-likelihood of a test sample under the GM-FLOW model is then used as the OOD indicator.

While density-based approaches have shown promising results, they often face challenges in accurately modeling the underlying distribution of ID samples, especially when the data is high-dimensional or has a complex structure. Additionally, these methods can be computationally expensive, as they require training generative models or estimating likelihood functions.

To address these limitations, recent research has explored hybrid approaches that combine density estimation with other techniques, such as feature-based methods [6] or self-supervised learning [54]. These hybrid methods aim to leverage the strengths of multiple approaches to improve the overall OOD detection performance.

Overall, density-based OOD detection methods provide a principled approach to identifying OOD samples by modeling the underlying distribution of ID data. While these methods have shown promising results, ongoing research is focused on addressing their limitations and developing more robust and efficient OOD detection solutions.

3.3 Distance-based Approaches

Distance-based approaches for out-of-distribution (OOD) detection leverage the distance between an input sample and the in-distribution (ID) data to identify OOD samples. The underlying assumption is that OOD samples will be relatively far from the ID data distribution compared to ID samples.

One of the earliest and most widely used distance-based OOD detection methods is the Mahalanobis distance [55]. The Mahalanobis distance measures the distance between a sample and the mean of the ID distribution, taking into account the covariance structure of the ID data. By computing the Mahalanobis distance of a test sample from the class-conditional means, the method can identify whether the sample is likely to belong to the ID distribution or not. The Mahalanobis distance-based approach has been widely adopted and extended in various OOD detection works [55].

Another family of distance-based OOD detection methods leverages nearest neighbor (NN) search [22]. The key idea is to measure the distance between a test sample and its nearest neighbors in the ID feature space. If a test sample is far from its nearest neighbors, it is likely to be an OOD sample. Compared to the Mahalanobis distance, NN-based methods do not assume a specific distribution of the ID data, making them more flexible and applicable to a wider range of scenarios. One challenge with NN-based methods is the computational cost of performing nearest neighbor search, especially for high-dimensional feature spaces. To address this, some works have proposed efficient approximate NN search algorithms [56].

Beyond the Mahalanobis distance and nearest neighbor-based approaches, researchers have explored other distance-based OOD detection techniques. For example, [30] proposes to model each ID class with multiple prototypes and use the distance between a test sample and the prototypes as the OOD score. [57] introduces a class relevance learning framework that leverages the inter-class relationships to enhance the distance-based OOD detection.

One common limitation of distance-based methods is that they rely on the assumption that OOD samples are far from the ID data distribution. However, in some cases, OOD samples may be close to the ID distribution, especially when the ID and OOD datasets share similar semantic content. To address this, some works have proposed to combine distance-based approaches with other techniques, such as density estimation [58] or classification-based approaches [59].

Overall, distance-based approaches provide a simple and intuitive way to identify OOD samples by measuring their proximity to the ID data distribution. While effective in many scenarios, they may struggle when the ID and OOD distributions overlap significantly. Combining distance-based methods with other techniques or developing more advanced distance metrics can help address these limitations and further improve OOD detection performance.

3.4 Reconstruction-based Approaches

Reconstruction-based approaches for out-of-distribution (OOD) detection leverage the premise that in-distribution (ID) samples can be reconstructed more accurately compared to OOD samples. These methods employ generative models or autoencoders to reconstruct the input and then use the reconstruction error as an indicator of whether the input is in-distribution or out-of-distribution.

One of the key advantages of reconstruction-based approaches is their ability to handle diverse types of OOD samples, as they do not make strong assumptions about the characteristics of OOD data. Instead, they rely on the underlying generative model's ability to accurately reconstruct ID samples, which is then used to identify OOD samples that deviate from this reconstruction.

A prominent example of a reconstruction-based OOD detection method is the Variational Autoencoder (VAE) [60; 61]. VAEs are generative models that learn a compact latent representation of the input data and can then be used to reconstruct the input. The key idea is that the reconstruction error, measured as the distance between the input and its reconstruction, can serve as an effective OOD indicator, as OOD samples will generally have higher reconstruction errors compared to ID samples.

Other reconstruction-based approaches, such as Generative Adversarial Networks (GANs) [62], have also been explored for OOD detection. In these methods, the generator component of the GAN is responsible for reconstructing the input, while the discriminator component is trained to distinguish between real ID samples and the reconstructed samples. The discriminator's output can then be used as an OOD score, as it will generally be lower for OOD samples compared to ID samples.

One of the key challenges in reconstruction-based OOD detection is ensuring that the generative model is sufficiently expressive and accurately captures the ID data distribution. If the model is unable to reconstruct ID samples well, it may lead to high reconstruction errors even for ID samples, resulting in poor OOD detection performance. To address this, some studies have explored ways to improve the generative model's reconstruction capabilities, such as by incorporating auxiliary tasks [63], using domain-specific data augmentation techniques [64], or employing multiple reconstruction heads [65].

Another line of research has focused on improving the OOD detection performance by leveraging the latent representations learned by the generative model, rather than just the reconstruction error. For example, [66] proposed a method that combines the reconstruction error with a confidence-based OOD score computed from the latent representations to achieve better OOD detection performance.

Furthermore, some studies have explored hybrid approaches that combine reconstruction-based methods with other OOD detection techniques, such as classification-based or density-based methods, to leverage the complementary strengths of different approaches [52]. By integrating multiple signals, these hybrid methods aim to achieve more robust and reliable OOD detection.

Overall, reconstruction-based approaches for OOD detection have shown promising results and offer the advantage of being able to handle diverse types of OOD samples. Ongoing research in this area focuses on improving the generative model's reconstruction capabilities, leveraging the learned latent representations, and combining reconstruction-based methods with other OOD detection techniques to enhance overall performance.

3.5 Feature-based Approaches

Feature-based approaches for out-of-distribution (OOD) detection leverage the learned features of the classification model to identify OOD samples. These approaches aim to capture the underlying representations and relationships within the in-distribution (ID) data, and then utilize these features to detect samples that deviate from the expected patterns.

One prominent line of feature-based OOD detection methods is based on contrastive learning. Contrastive learning techniques learn representations by encouraging the model to bring nearby (positive) samples closer in the feature space while pushing away distant (negative) samples. This results in more compact and discriminative feature representations, which can then be used to identify OOD samples that are distant from the ID data manifold.

For example, the paper "Shifting Transformation Learning for Out-of-Distribution Detection" [67] proposes a framework that leverages a shifting transformation learning setting to learn multiple shifted representations of the training set for improved OOD detection. The authors argue that the choice of shifting transformations and pretext tasks in self-supervised representation learning techniques often depends on the in-domain distribution, and they introduce a simple mechanism for automatically selecting the transformations and modulating their effect on representation learning without requiring any OOD training samples.

Another feature-based approach is the prototype-based learning, which models each class with multiple prototypes to capture the natural diversities within the data. The paper "Learning with Mixture of Prototypes for Out-of-Distribution Detection" [30] introduces PrototypicAl Learning with a Mixture of prototypes (PALM), which automatically identifies and dynamically updates prototypes, assigning each sample to a subset of prototypes via reciprocal neighbor soft assignment weights. PALM optimizes a maximum likelihood estimation (MLE) loss to encourage the sample embeddings to be compact around the associated prototypes, as well as a contrastive loss on all prototypes to enhance intra-class compactness and inter-class discrimination at the prototype level. The authors demonstrate that this approach, which models each class with multiple prototypes, can significantly improve OOD detection performance compared to methods that rely on a single prototype per class.

Furthermore, the paper "Class Relevance Learning For Out-of-distribution Detection" [57] presents an innovative class relevance learning method tailored for OOD detection. The proposed framework establishes a comprehensive class relevance learning approach, strategically harnessing

the intricate interclass relationships within the OOD pipeline. The authors show that this framework can significantly augment OOD detection capabilities compared to state-of-the-art alternatives.

In addition to contrastive learning and prototype-based methods, other feature-based approaches have also been explored. The paper "Boosting Out-of-distribution Detection with Typical Features" [68] introduces the concept of "typical features", which are the high-probability region of the deep model's feature space. The authors propose to rectify the feature into its typical set and calculate the OOD score with the typical features to achieve reliable uncertainty estimation. They demonstrate that this feature rectification approach can be used as a plug-and-play module with various OOD scores to boost the OOD detection performance.

The key advantage of feature-based OOD detection methods is their ability to capture the underlying structure and relationships within the ID data, which can provide more informative cues for identifying OOD samples. By leveraging the learned representations, these approaches can often achieve robust OOD detection performance without relying on explicit modeling of the ID or OOD data distributions. However, the effectiveness of feature-based methods can be sensitive to the quality and relevance of the learned representations, which may depend on the specific architecture, training dataset, and task at hand.

4. Benchmark Datasets and Evaluation Protocols

4.1 Commonly Used Benchmark Datasets

Out-of-distribution (OOD) detection methods are typically evaluated on a variety of commonly used benchmark datasets, such as CIFAR-10, CIFAR-100, ImageNet, and LSUN. These standardized datasets provide a platform for assessing the performance of OOD detection algorithms and enable fair comparisons across different methods.

The CIFAR-10 and CIFAR-100 datasets are widely used in the field of OOD detection [5]. CIFAR-10 consists of 60,000 low-resolution (32x32) color images belonging to 10 classes, while CIFAR-100 has 100 classes with 600 images per class. However, these datasets have a limited number of classes and relatively simple visual structures, which may not capture the complexity of real-world OOD samples that an OOD detection system is likely to encounter in deployment [6; 69].

In contrast, the ImageNet dataset is a large-scale benchmark that provides a more diverse and challenging testbed for evaluating OOD detection methods [10]. It consists of over 14 million high-resolution images belonging to 1,000 classes, offering a broader range of visual and semantic information. The LSUN dataset, which includes over 10 million images across 10 scene categories, is another popular choice for assessing OOD detection performance [22].

While these benchmark datasets have been widely adopted, they also have limitations. The images within these datasets may not capture the full range of variations and complexity present in real-world scenarios, such as sensor degradation, environmental changes, or adversarial

perturbations [4]. Additionally, the OOD samples in these datasets are often manually curated, which can introduce bias and may not accurately reflect the distribution of OOD samples that a deployed model might encounter [70].

To address these limitations, researchers have proposed more diverse and challenging benchmark datasets for OOD detection. For example, the OpenOOD framework [5] includes a comprehensive set of benchmark datasets spanning different domains, such as medical imaging, satellite imagery, and natural language processing. This unified and well-structured benchmark aims to drive the field of OOD detection forward by enabling fair comparisons and identifying the strengths and weaknesses of various OOD detection methods.

Moreover, recent studies have highlighted the importance of evaluating OOD detection methods on a broader range of datasets, including those with more realistic and diverse OOD samples [4]. This includes datasets with natural distribution shifts, such as weather changes or sensor degradation, as well as datasets with adversarial perturbations that can fool OOD detectors.

In summary, the commonly used benchmark datasets for OOD detection, while providing a standardized platform, also have limitations in terms of the diversity and complexity of OOD samples. Researchers are actively working to develop more comprehensive and realistic benchmark datasets to better evaluate the real-world performance of OOD detection systems, which is crucial for their successful deployment in practical applications.

4.2 Evaluation Protocols for OOD Detection

Evaluating the performance of out-of-distribution (OOD) detection methods is a crucial aspect in the development and deployment of reliable machine learning systems. The research community has adopted several standard evaluation protocols and metrics to assess the effectiveness of OOD detection approaches.

One of the most widely used metrics is the Area Under the Receiver Operating Characteristic (AUROC) curve, which measures the overall ability of the OOD detector to distinguish between in-distribution (ID) and OOD samples [12; 71; 72]. A value of 1 indicates perfect separation, while 0.5 represents random guessing.

Another commonly used metric is the False Positive Rate at 95% True Positive Rate (FPR95) [12; 73]. FPR95 quantifies the percentage of OOD samples that are falsely classified as ID when the true positive rate (the percentage of ID samples correctly identified) is 95%. A lower FPR95 indicates better OOD detection performance.

While AUROC and FPR95 are widely adopted, they have their own strengths and weaknesses. AUROC provides an overall assessment, but may not fully capture the relative importance of true positives and false positives. FPR95, on the other hand, focuses on the high-confidence region, but may not reflect the performance across the entire operating range [73].

To address these limitations, researchers have proposed alternative metrics, such as Normalized Accuracy (NA) and Youden's J-statistic [35]. NA combines the classification accuracy on ID samples and the detection rate on OOD samples, offering a more holistic measure. Youden's J-statistic considers both the true positive rate and the true negative rate, providing a balanced assessment of the OOD detection capability.

Recent efforts have also aimed to develop evaluation protocols that better reflect real-world deployment scenarios. For instance, the Semantic Shift Benchmark (SSB) [8] focuses on detecting semantic novelty, rather than just differences in datasets, to capture practical challenges faced by OOD detectors.

Furthermore, researchers have emphasized the importance of evaluating OOD detection methods on diverse datasets and scenarios [12; 74]. Existing benchmarks often use a limited set of OOD datasets, which may not capture the full spectrum of distribution shifts encountered in practice. More comprehensive evaluation protocols that include a wider range of OOD datasets have been proposed to better guide the development of robust OOD detection methods.

4.3 Limitations of Existing Benchmarks

While the current benchmarks for out-of-distribution (OOD) detection have served as valuable tools for evaluating the performance of various methods, they suffer from several limitations that may lead to biased or unrealistic assessments.

One of the primary limitations is the lack of diversity in the OOD datasets used for evaluation. Many studies have highlighted the tendency of OOD detection methods to perform well on datasets that exhibit a clear semantic shift from the in-distribution (ID) data, such as CIFAR-10 versus CIFAR-100 or ImageNet [6; 11]. However, in real-world applications, OOD data may not necessarily exhibit such a stark contrast and could instead exhibit more subtle distribution shifts, such as changes in image resolution, lighting conditions, or object viewpoints.

Furthermore, the OOD datasets used in many benchmarks are often constructed by selecting images from entirely different datasets, such as LSUN or Textures, which may not accurately reflect the types of OOD data that a deployed model is likely to encounter. This lack of diversity in the OOD datasets can lead to the development of OOD detection methods that are overly specialized and fail to generalize to more realistic OOD scenarios [11].

Another limitation of existing benchmarks is the potential for dataset biases and artifacts to influence the performance of OOD detection methods. Many datasets used for OOD detection, such as CIFAR-10 and CIFAR-100, have been extensively studied and may contain subtle statistical patterns or correlations that are exploited by OOD detection methods, leading to inflated performance [75]. This can result in OOD detection methods that perform well on the benchmark but may not generalize to real-world scenarios where the dataset biases and artifacts are different.

The evaluation protocols used in many OOD detection benchmarks also present limitations. Commonly used metrics, such as the area under the receiver operating characteristic (AUROC) curve and the false positive rate at 95% true positive rate (FPR95), may not provide a comprehensive assessment of the performance of OOD detection methods. These metrics focus on the overall separation between in-distribution and out-of-distribution data, but they do not capture the nuances of how OOD detection methods handle different types of distribution shifts or the trade-offs between false positive and false negative rates [13].

Additionally, the majority of OOD detection benchmarks have focused on image classification tasks, while there is a need for more diverse evaluation scenarios, including natural language processing, time series analysis, and other domains where OOD detection is equally crucial [3].

To address these limitations, researchers have proposed more comprehensive and realistic evaluation frameworks for OOD detection. For example, [13] introduced the CIFAR-10-R, CIFAR-100-R, and ImageNet-30-R datasets, which aim to capture more realistic distribution shifts, and the Generalizability Score (GS) metric, which measures the ability of OOD detection methods to generalize across different types of distribution shifts.

Furthermore, [11] proposed the ImageNet-OOD dataset, which decouples semantic shift and covariate shift, providing a more targeted evaluation of OOD detection methods' performance on these different types of distribution shifts.

Overall, the limitations of existing benchmarks for OOD detection highlight the need for more diverse, realistic, and comprehensive evaluation frameworks that can better assess the performance of OOD detection methods in real-world scenarios. By addressing these limitations, the research community can develop more robust and reliable OOD detection techniques that can be safely deployed in critical applications.

4.4 Towards More Comprehensive and Realistic Evaluation

Current benchmark datasets and evaluation protocols for out-of-distribution (OOD) detection have significant limitations that hinder the development of robust and practical OOD detection methods. Addressing these limitations is crucial for advancing the field and enabling the safe deployment of machine learning models in real-world applications.

One of the key challenges is the lack of data diversity in existing benchmark datasets. Many commonly used datasets, such as CIFAR-10, CIFAR-100, and ImageNet, have relatively narrow data distributions, often focusing on specific object categories or visual domains [13]. This limited diversity does not reflect the wide range of inputs that models may encounter in real-world scenarios, where data can come from diverse sources, settings, and modalities.

To address this issue, researchers have proposed developing more comprehensive and diverse benchmark datasets that better capture the complexity and heterogeneity of real-world data. For example, the DAWN dataset [76] includes images with various weather conditions, and the OOD-CV dataset [77] includes images with different poses, shapes, textures, and environmental contexts. These datasets aim to challenge OOD detection models by introducing more realistic distribution shifts and nuisance factors.

Another important aspect of realistic evaluation is accounting for distribution shifts and dynamic changes in data distributions over time. Many existing benchmark datasets assume a static and pre-defined separation between in-distribution and out-of-distribution data, which may not reflect the more fluid and continuous nature of distribution shifts in real-world scenarios [16]. To address this, researchers have proposed evaluation setups that simulate distribution shifts, such as the Generalized Out-of-Distribution Detection (OpenOOD) framework [5] and the Continuously Adaptive Out-of-Distribution (CAOOD) detection setting [16].

Additionally, current evaluation metrics, such as Area Under the Receiver Operating Characteristic (AUROC) and False Positive Rate at 95% True Positive Rate (FPR95), may not fully capture the practical requirements of OOD detection in real-world applications [14]. These metrics often focus on binary classification performance and may not adequately assess the separation between in-distribution and out-of-distribution data, especially when there is significant overlap between the two distributions. Researchers have proposed alternative evaluation metrics, such as the Area Under the Threshold Curve (AUTC) [14], which better capture the quality of the separation between in-distribution and out-of-distribution data.

Furthermore, current benchmark datasets and evaluation protocols often lack consideration for the practical constraints and requirements of real-world applications. For instance, in safety-critical systems like autonomous vehicles or medical diagnosis, the cost of false positives and false negatives can be vastly different, and the evaluation should reflect these differential consequences [78]. Researchers have proposed addressing this by incorporating application-specific requirements and constraints into the evaluation protocols, such as the need for low false positive rates or the ability to handle diverse types of out-of-distribution data.

In summary, developing more comprehensive and realistic benchmark datasets and evaluation protocols for OOD detection is essential to advancing the field and ensuring the safe deployment of machine learning models in real-world applications. This includes incorporating data diversity, distribution shifts, and practical constraints into the evaluation framework, as well as exploring more suitable evaluation metrics that better align with the requirements of specific application domains. By addressing these limitations, the research community can drive the development of OOD detection methods that are more robust, reliable, and suitable for deployment in complex, dynamic, and safety-critical environments.

5. Practical Considerations and Applications

5.1 Practical Challenges in Deploying OOD Detection Methods

Deploying out-of-distribution (OOD) detection methods in real-world applications poses several practical challenges that must be addressed to ensure their reliable and effective performance. One key challenge is the need for data diversity [69; 79]. OOD detection methods often rely on the assumption that the training data, known as in-distribution (ID) data, adequately represents the expected input distribution. However, in practice, the real-world data encountered during deployment can be significantly more diverse and complex than the training data, leading to the emergence of various types of OOD samples [6].

To handle this challenge, OOD detection methods must be designed to address a wide range of OOD samples, including those that are semantically similar to the ID data but still lie outside the training distribution [10; 70]. This requires a thorough understanding of the potential sources of distribution shifts, such as covariate shifts and semantic shifts, and the development of robust techniques that can effectively detect these diverse OOD samples [10].

Another practical challenge is the need for robustness to distribution shifts over time [16; 80]. In many real-world applications, the data distribution may change dynamically due to factors such as sensor degradation, environmental changes, or the introduction of new types of data. Traditional OOD detection methods that rely on a static decision boundary may struggle to maintain their performance in the face of these distribution shifts. To address this challenge, OOD detection methods must be designed to adapt to evolving distributions, either by continuously updating the decision boundary or by leveraging online data to refine the OOD detection mechanism [16; 80].

Computational efficiency is another crucial practical consideration when deploying OOD detection methods, particularly in resource-constrained environments such as embedded systems or real-time applications [1]. Many existing OOD detection methods, especially those based on deep learning, can be computationally intensive, requiring significant processing power and memory resources. To address this challenge, OOD detection methods must be optimized for efficient implementation, potentially through techniques such as model compression, quantization, or the use of lightweight neural network architectures [1; 81].

Furthermore, the integration of OOD detection capabilities with the primary task of the machine learning model is crucial for practical deployment [4]. OOD detection should not compromise the performance of the main task, and the two components should work harmoniously to ensure the overall reliability and safety of the system. Approaches that can jointly optimize for both in-distribution performance and OOD detection, or that can seamlessly incorporate OOD detection as a post-processing step, are particularly desirable for practical applications [22; 82].

5.2 Out-of-Distribution Detection in Time-Series Domain

Out-of-distribution (OOD) detection in the time-series domain poses unique challenges compared to the image domain. Unlike static images, time-series data, such as stock prices, sensor measurements, or audio recordings, exhibit inherent temporal dependencies and dynamic patterns that evolve over time. This temporal aspect introduces additional complexities that need to be addressed when developing effective OOD detection methods for time-series data.

One of the key challenges in time-series OOD detection is the presence of seasonal or periodic patterns within the data. These recurring patterns can be mistaken for anomalous behavior, leading to high false positive rates in OOD detection [6]. To address this issue, researchers have proposed the Seasonal Ratio Scoring (SRS) approach, which aims to distinguish seasonal variations from genuine OOD events.

The SRS method, as described in [6], leverages the intuition that OOD samples will exhibit a significantly different ratio between the current value and the corresponding seasonal value, compared to in-distribution samples. By computing this ratio and analyzing its distribution, the SRS approach can effectively identify OOD instances that deviate from the expected seasonal pattern. This technique is particularly well-suited for time-series data, as it considers the temporal dynamics and seasonal behaviors inherent in the data, allowing it to better distinguish between normal fluctuations and truly anomalous events, resulting in improved OOD detection performance.

Furthermore, the SRS approach can be used in an unsupervised setting, which is particularly advantageous for time-series data, where obtaining labeled OOD samples may be challenging or even infeasible. By solely relying on the inherent seasonal patterns within the in-distribution data, the SRS method can identify OOD instances without the need for explicit labeling of the training data.

In addition to the SRS approach, researchers have explored other techniques for OOD detection in time-series data. For example, some studies have proposed the use of deep learning-based models, such as recurrent neural networks (RNNs) or temporal convolutional networks (TCNs), to capture the complex temporal dependencies within the data and identify anomalous patterns [6]. These models can learn robust representations of the in-distribution data and leverage them for effective OOD detection.

Another strategy involves the integration of multiple OOD detection methods, each targeting different aspects of the time-series data. By combining complementary approaches, such as seasonal-aware techniques and deep learning-based models, researchers have demonstrated improved performance in identifying diverse types of OOD events in time-series data [6].

In summary, the unique challenges posed by time-series data, such as the presence of seasonal patterns and dynamic temporal dependencies, require specialized OOD detection methods. The Seasonal Ratio Scoring (SRS) approach, as described in [6], is a notable example of an effective technique that addresses these challenges by leveraging the inherent

seasonal characteristics of the data. As the field of time-series OOD detection continues to evolve, a combination of domain-specific techniques and advancements in deep learning may lead to further improvements in the reliability and robustness of OOD detection for real-world time-series applications.

5.3 Unified Out-of-Distribution Detection: A Model-Specific Perspective

Out-of-distribution (OOD) detection in the time-series domain poses unique challenges compared to the image domain. Unlike static images, time-series data, such as stock prices, sensor measurements, or audio recordings, exhibit inherent temporal dependencies and dynamic patterns that evolve over time. This temporal aspect introduces additional complexities that need to be addressed when developing effective OOD detection methods for time-series data.

One of the key challenges in time-series OOD detection is the presence of seasonal or periodic patterns within the data. These recurring patterns can be mistaken for anomalous behavior, leading to high false positive rates in OOD detection [6]. To address this issue, researchers have proposed the Seasonal Ratio Scoring (SRS) approach, which aims to distinguish seasonal variations from genuine OOD events.

The SRS method, as described in [6], leverages the intuition that OOD samples will exhibit a significantly different ratio between the current value and the corresponding seasonal value, compared to in-distribution samples. By computing this ratio and analyzing its distribution, the SRS approach can effectively identify OOD instances that deviate from the expected seasonal pattern.

In contrast to traditional OOD detection methods that rely on static features or statistics, the SRS approach is particularly well-suited for time-series data as it takes into account the temporal dynamics and seasonal behaviors inherent in the data. This allows the method to better distinguish between normal fluctuations and truly anomalous events, leading to improved OOD detection performance in the time-series domain.

Furthermore, the SRS approach can be used in an unsupervised setting, which is particularly advantageous for time-series data, where obtaining labeled OOD samples may be challenging or even infeasible. By solely relying on the inherent seasonal patterns within the in-distribution data, the SRS method can identify OOD instances without the need for explicit labeling of the training data.

In addition to the SRS approach, researchers have explored other techniques for OOD detection in time-series data. For example, some studies have proposed the use of deep learning-based models, such as recurrent neural networks (RNNs) or temporal convolutional networks (TCNs), to capture the complex temporal dependencies within the data and identify anomalous patterns [6]. These models can learn robust representations of the in-distribution data and leverage them for effective OOD detection.

Another strategy involves the integration of multiple OOD detection methods, each targeting different aspects of the time-series data. By combining complementary approaches, such as seasonal-aware techniques and deep learning-based models, researchers have demonstrated improved performance in identifying diverse types of OOD events in time-series data [6].

In summary, the unique challenges posed by time-series data, such as the presence of seasonal patterns and dynamic temporal dependencies, require specialized OOD detection methods. The Seasonal Ratio Scoring (SRS) approach, as described in [6], is a notable example of an effective technique that addresses these challenges by leveraging the inherent seasonal characteristics of the data. As the field of time-series OOD detection continues to evolve, a combination of domain-specific techniques and advancements in deep learning may lead to further improvements in the reliability and robustness of OOD detection for real-world time-series applications.

5.4 Model-free Test Time Adaptation for Out-Of-Distribution Detection

Out-of-distribution (OOD) detection is a critical task in ensuring the reliability and safety of machine learning models in real-world applications. Conventional OOD detection methods often rely on a fixed decision criterion, learned from a given in-distribution (ID) dataset, to identify OOD samples during deployment [21]. However, this approach can be limiting, as it may not adapt well to the evolving distributions encountered in dynamic environments.

To address this challenge, researchers have recently explored the idea of leveraging test-time adaptation to enhance the adaptability of OOD detectors to changing data distributions [21]. The key insight behind this approach is that by utilizing the information available in online test samples, OOD detectors can be adapted during the deployment phase, enabling them to better cope with distribution shifts.

One such method that exemplifies this approach is the Non-Parametric Test Time Adaptation framework for Out-Of-Distribution Detection (ATTA) [21]. ATTA aims to enhance the adaptability of OOD detectors by incorporating online test samples into the decision-making process, rather than relying solely on the fixed model learned from the training data.

The core idea behind ATTA is to leverage the power of non-parametric techniques to adapt the OOD detection model during the test phase. Unlike parametric methods, which rely on pre-defined functional forms, non-parametric approaches are more flexible and can better capture the complex and evolving relationships within the data [21].

In the ATTA framework, the OOD detection model is initially trained on the available ID dataset. During the test phase, however, the model is continuously adapted by incorporating the information from the incoming test samples. This adaptation process is performed in a non-parametric manner, allowing the model to adapt to the changing data distributions without the need for extensive retraining or fine-tuning.

The key advantage of ATTA is its ability to leverage the information from the online test samples to enhance the OOD detection performance, even in the presence of distribution shifts. By continuously adapting the model, ATTA can better distinguish between ID and OOD samples, particularly when the ID and OOD distributions overlap significantly [21].

Extensive experiments conducted on various OOD detection benchmarks have demonstrated the effectiveness of the ATTA framework. Compared to state-of-the-art OOD detection methods, ATTA has been shown to significantly improve the performance, reducing the false positive rate (FPR95) by up to 23.23% on the CIFAR-10 benchmark and 38% on the ImageNet-1k benchmark [21].

The success of ATTA can be attributed to its ability to adapt to the changing data distributions, which is a crucial requirement for the reliable deployment of machine learning models in real-world scenarios. By leveraging non-parametric techniques and online test samples, ATTA can maintain a high level of adaptability, ensuring that the OOD detection model can effectively handle the diverse distribution shifts that may occur during the deployment phase.

Overall, the ATTA framework highlights the importance of incorporating test-time adaptation strategies for enhancing the robustness and reliability of OOD detectors in dynamic environments. As machine learning models are increasingly deployed in safety-critical applications, the need for such adaptive and flexible OOD detection approaches becomes more pressing, paving the way for further research and advancements in this field.

5.5 Learning with Mixture of Prototypes for Out-of-Distribution Detection

Out-of-distribution (OOD) detection is a critical task for ensuring the reliability and safety of machine learning models deployed in real-world applications. Existing distance-based OOD detection methods often rely on learning a representation that models each class with a single prototype or centroid, which can be an oversimplified assumption that overlooks the natural diversity within the data. This limitation can lead to inadequate modeling of realistic data and suboptimal OOD detection performance.

To address this issue, the authors of [30] propose PrototypicAl Learning with a Mixture of prototypes (PALM), a novel method that models each class with multiple prototypes to capture the sample diversities and learn more faithful and compact sample embeddings to enhance OOD detection. Unlike conventional approaches that rely on a fixed decision criterion learned from the training data, PALM leverages the flexibility of non-parametric techniques to continuously adapt the OOD detection model during the test phase, as described in the previous subsection.

The key idea behind PALM is to encourage the sample embeddings to be compact around the associated prototypes, as well as to enhance the intra-class compactness and inter-class discrimination at the prototype level. Specifically, PALM optimizes a maximum likelihood estimation (MLE) loss to encourage the sample embeddings to be compact around the

associated prototypes, and a contrastive loss on all prototypes to enhance the intra-class compactness and inter-class discrimination.

Moreover, the automatic estimation of prototypes in PALM enables the approach to be extended to the challenging OOD detection task with unlabeled in-distribution (ID) data, which is a common scenario in real-world applications. This is a significant advantage over methods that rely on labeled ID data or require a predefined number of prototypes per class.

The authors evaluate PALM extensively on various benchmarks, including the challenging CIFAR-100 dataset, and demonstrate its superiority over state-of-the-art OOD detection methods. On CIFAR-100, PALM achieves a state-of-the-art average AUROC performance of 93.82, outperforming previous methods by a wide margin.

One of the key strengths of PALM is its ability to learn a more faithful and compact representation of the data by modeling each class with multiple prototypes. This is in contrast to the oversimplified assumption of a single prototype per class, which can lead to suboptimal performance in real-world scenarios where the data exhibits significant diversity within each class. Furthermore, the automatic estimation of prototypes in PALM allows the method to be easily adapted to different datasets and tasks, without the need for manual tuning of the number of prototypes per class. This flexibility and adaptability are particularly important in the context of OOD detection, where the data distribution can vary widely across different applications and deployment scenarios.

In conclusion, the PALM method proposed in [30] represents a significant advancement in the field of OOD detection. By modeling each class with a mixture of prototypes, PALM is able to learn a more faithful and compact representation of the data, leading to improved OOD detection performance. The authors' comprehensive evaluation and the method's ability to handle unlabeled ID data further underscore the practical relevance and potential impact of PALM in real-world applications.

6. Emerging Trends and Future Directions

6.1 Dealing with Distribution Shifts

Out-of-distribution (OOD) detection is a crucial capability for the safe deployment of machine learning models in real-world environments, where the data encountered during deployment may differ significantly from the training distribution. A key challenge in OOD detection is the ability to handle distributional shifts, which can take various forms, such as covariate shifts and semantic shifts.

Covariate shifts refer to changes in the input data distribution, while semantic shifts involve changes in the underlying concepts or classes represented in the data. These shifts can occur due to a wide range of factors, such as changes in the data collection process, sensor degradation, environmental changes, or the introduction of new data sources. Handling such distribution shifts is essential for maintaining the reliability and robustness of OOD detectors, as models trained on a

static distribution may fail to generalize to the dynamic and evolving nature of real-world data.

Recent research has highlighted the limitations of existing OOD detection methods in dealing with distribution shifts. For example, [10] has shown that many OOD detection approaches focus solely on semantic shifts, ignoring the potential impact of covariate shifts. This can lead to the failure of these methods in real-world scenarios where both types of shifts may occur simultaneously.

To address this challenge, researchers have proposed various strategies for developing OOD detectors that can adapt to dynamic distributions. One approach is to leverage online or continual learning techniques, where the OOD detector is constantly updated with new data, allowing it to adapt to changing distributions [16]. These methods aim to strike a balance between maintaining performance on the in-distribution data and enhancing the detector's ability to handle distribution shifts.

Another direction is to explore unsupervised or self-supervised learning techniques that can learn robust feature representations resilient to distribution shifts [83]. By learning representations that capture the underlying structure of the data, rather than relying solely on surface-level features, these approaches can potentially improve the generalization of OOD detectors to a broader range of distribution shifts.

Additionally, researchers have investigated the use of adversarial training and data augmentation techniques to enhance the robustness of OOD detectors to distribution shifts [84]. By exposing the models to a diverse range of distributional variations during training, these methods aim to improve the detectors' ability to handle unseen shifts during deployment.

Furthermore, the development of meta-learning and few-shot learning approaches for OOD detection [16] holds promise for rapidly adapting to new distribution shifts with limited data or supervision. These techniques could enable OOD detectors to quickly adjust to emerging distribution changes, rather than relying on costly retraining or fine-tuning.

In summary, the challenge of handling distribution shifts is a critical and ongoing concern in the field of OOD detection. Addressing this challenge requires the development of adaptive, robust, and versatile OOD detectors that can maintain reliable performance in the face of dynamic and evolving data distributions. Continued research in this direction, drawing insights from various machine learning paradigms, will be crucial for the safe and widespread deployment of machine learning models in real-world applications.

6.2 Leveraging Unlabeled Data

The lack of out-of-distribution (OOD) data during training is a key challenge in OOD detection, as traditional methods often rely on carefully curated in-distribution data. However, the abundance of

unlabeled data in real-world scenarios has led researchers to explore ways to leverage this unlabeled information to enhance OOD detection performance.

One promising direction is the use of self-supervised learning, which aims to learn meaningful representations from unlabeled data by designing pretext tasks [85; 72]. The self-supervised features captured by these methods often encode more general and robust patterns that can be useful for distinguishing in-distribution and out-of-distribution samples, thereby improving the OOD detection capabilities of the model.

Another approach is to leverage unlabeled data through sample repairing techniques, where the goal is to generate synthetic in-distribution samples from the unlabeled data [37]. By "repairing" the missing in-distribution samples, these methods can enhance the model's ability to differentiate between in-distribution and OOD samples.

Additionally, researchers have explored methods that aim to separate candidate outliers from the unlabeled data [8]. By identifying potential OOD samples within the unlabeled data, these techniques can then use the remaining, more reliable unlabeled data to improve the OOD detection performance.

The use of unlabeled data has shown promising results in enhancing OOD detection. For example, [85] demonstrated that self-supervised learning can significantly boost the performance of OOD detectors, particularly for high-dimensional data like images. Similarly, [37] showed that sample repairing techniques can effectively enhance one-class classifiers for novelty detection, outperforming traditional methods.

However, the effective utilization of unlabeled data for OOD detection remains an active area of research, with several challenges and open questions. The reliability and robustness of these methods can be affected by the quality and diversity of the unlabeled samples, as well as the potential presence of OOD samples within the unlabeled data. Additionally, the integration of these techniques with other OOD detection approaches, such as density-based or distance-based methods, can lead to further performance improvements and provide a more comprehensive understanding of the OOD detection problem.

As the availability of unlabeled data continues to grow, exploring novel ways to effectively leverage this abundance of information to enhance OOD detection capabilities is a promising direction for future research. By combining the strengths of self-supervised learning, sample repairing, and outlier separation techniques, researchers can work towards more robust and generalizable OOD detection solutions that can be deployed in a wide range of real-world applications.

6.3 Integrating OOD Detection with Other ML Tasks

Out-of-distribution (OOD) detection has emerged as a critical task in machine learning, ensuring the reliable and safe deployment of models in real-world applications. While significant progress has been made in developing effective OOD detection methods, there is a growing need to

integrate OOD detection with other machine learning tasks to achieve a more holistic and comprehensive approach to handling distribution shifts and model uncertainty.

One promising area of integration is with open-set recognition, which aims to classify known classes while also identifying unknown classes during inference [86]. Open-set recognition and OOD detection share the common goal of detecting samples that do not belong to the training distribution, but they approach the problem from different perspectives. Open-set recognition focuses on identifying unknown classes, while OOD detection focuses on identifying samples that are outside the distribution of the training data, regardless of the class. By integrating these two tasks, researchers can develop models that not only detect OOD samples but also identify the specific class or category of the OOD sample, providing more informative and actionable output to end-users.

Another area of potential integration is with anomaly detection. Anomaly detection and OOD detection are closely related, as both aim to identify samples that deviate from the expected or normal distribution. However, the two tasks differ in their specific objectives and application domains. Anomaly detection is often used in industrial and security applications to identify unusual or suspicious patterns, while OOD detection is more commonly used in machine learning applications to ensure the reliability of model predictions [9]. By integrating these two tasks, researchers can develop models that can simultaneously detect anomalies and OOD samples, providing a more comprehensive solution for ensuring the safety and reliability of machine learning systems.

A third area of potential integration is with few-shot OOD detection, which aims to detect OOD samples when only a few examples of the OOD distribution are available during training [44]. This is a particularly relevant scenario in real-world applications, where the OOD distribution may be difficult or costly to obtain. By integrating few-shot OOD detection with other machine learning tasks, such as transfer learning or meta-learning, researchers can develop models that can adapt to new OOD distributions with limited training data, further enhancing the robustness and versatility of OOD detection systems.

The benefits of a more holistic approach to OOD detection are manifold. By integrating OOD detection with other machine learning tasks, researchers can develop models that can better understand and reason about the broader context of the data, leading to more accurate and reliable predictions. Additionally, a more integrated approach can help to reduce the computational and data requirements of OOD detection, as the model can leverage synergies between different tasks to improve overall performance. Furthermore, a holistic approach to OOD detection can lead to more interpretable and explainable models, as the integration of different tasks can provide additional insights into the decision-making process of the model.

6.4 Developing Interpretable and Robust OOD Detectors

The need for developing interpretable and robust out-of-distribution (OOD) detectors has become increasingly apparent in the field of machine learning. As OOD detection becomes more critical for ensuring the reliability and safety of AI systems, particularly in high-stakes applications, there is a growing recognition that existing OOD detection methods may not be adequate.

One key challenge is the lack of interpretability in many OOD detection methods. Most current approaches treat the OOD detection task as a black-box problem, focusing solely on optimizing the detection performance without providing insights into the underlying decision-making process [31]. This makes it difficult to understand why a particular input is identified as OOD, which can hinder the trustworthiness and transparency of the AI system.

To address this issue, recent research has explored the development of interpretable OOD detectors. For instance, [31] proposed a framework that provides concept-based explanations for OOD detectors, allowing users to understand the key factors contributing to the detection results. By identifying the high-level concepts that are most influential in the OOD detection process, this approach can offer valuable insights and improve the interpretability of the system.

In addition to interpretability, the robustness of OOD detectors to various distribution shifts and adversarial attacks is another crucial concern. As highlighted in [87; 68], the vulnerability of OOD detectors to distribution shifts can significantly degrade their performance, which is particularly problematic in real-world scenarios where the deployment environment is often subject to dynamic and unpredictable changes.

To enhance the robustness of OOD detectors, researchers have explored several promising directions. [87] investigated the impact of the backbone architecture on the OOD robustness and proposed techniques to minimize the distortion of the backbone features during fine-tuning, effectively preserving the generalizability of the model. [68] introduced a feature rectification approach that projects the input features onto their typical set, improving the robustness of the OOD detector to distribution shifts.

Furthermore, the development of OOD detectors that are robust to adversarial attacks has also gained attention. [88] proposed a novel evaluation protocol that utilizes adversarial generation of OOD samples to assess the robustness of OOD detectors, revealing previously overlooked weaknesses in existing methods.

In summary, the development of interpretable and robust OOD detectors is a critical area of research that will significantly enhance the reliability and safety of AI systems in real-world applications. By providing insights into the decision-making process and ensuring the resilience of OOD detectors to various distribution shifts and adversarial attacks, these advancements can pave the way for the widespread deployment of trustworthy and dependable AI systems.

References

[1] Design Methodology for Deep Out-of-Distribution Detectors in Real-Time Cyber-Physical Systems

[2] A Benchmark of Medical Out of Distribution Detection

[3] A Survey on Out-of-Distribution Detection in NLP

[4] A Critical Evaluation of Open-World Machine Learning

[5] OpenOOD Benchmarking Generalized Out-of-Distribution Detection

[6] Generalized Out-of-Distribution Detection A Survey

[7] Classification-Based Anomaly Detection for General Data

[8] Open-Set Recognition a Good Closed-Set Classifier is All You Need

[9] A Unified Survey on Anomaly, Novelty, Open-Set, and Out-of-Distribution Detection Solutions and Future Challenges

[10] Unified Out-Of-Distribution Detection A Model-Specific Perspective

[11] ImageNet-OOD Deciphering Modern Out-of-Distribution Detection Algorithms

[12] Practical Evaluation of Out-of-Distribution Detection Methods for Image Classification

[13] Towards Realistic Out-of-Distribution Detection A Novel Evaluation Framework for Improving Generalization in OOD Detection

[14] Beyond AUROC & co. for evaluating out-of-distribution detection performance

[15] Unsupervised Evaluation of Out-of-distribution Detection A Data-centric Perspective

[16] Meta OOD Learning for Continuously Adaptive OOD Detection

[17] Towards Effective Semantic OOD Detection in Unseen Domains A Domain Generalization Perspective

[18] Improving Variational Autoencoder based Out-of-Distribution Detection for Embedded Real-time Applications

[19] CODiT Conformal Out-of-Distribution Detection in Time-Series Data

[20] WiP Abstract Robust Out-of-distribution Motion Detection and Localization in Autonomous CPS

[21] Model-free Test Time Adaptation for Out-Of-Distribution Detection

[22] Out-of-Distribution Detection with Deep Nearest Neighbors

[23] Is Fine-tuning Needed Pre-trained Language Models Are Near Perfect for Out-of-Domain Detection

[24] Mind the Backbone Minimizing Backbone Distortion for Robust Object Detection

[25] MIM-OOD Generative Masked Image Modelling for Out-of-Distribution Detection in Medical Images

[26] Types of Out-of-Distribution Texts and How to Detect Them

[27] Cadence A Practical Time-series Partitioning Algorithm for Unlabeled IoT Sensor Streams

[28] Anomaly Detection under Distribution Shift

[29] How Does Unlabeled Data Provably Help Out-of-Distribution Detection

[30] Learning with Mixture of Prototypes for Out-of-Distribution Detection

[31] Concept-based Explanations for Out-Of-Distribution Detectors

[32] Out-of-Distribution Data An Acquaintance of Adversarial Examples -- A Survey

[33] A Survey on Explainable Anomaly Detection

[34] Continual Novelty Detection

[35] A Survey on Open Set Recognition

[36] Deep Learning for Anomaly Detection A Review

[37] Adversarially Learned One-Class Classifier for Novelty Detection

[38] A noisy elephant in the room Is your out-of-distribution detector robust to label noise

[39] DOODLER Determining Out-Of-Distribution Likelihood from Encoder Reconstructions

[40] Contrastive Training for Improved Out-of-Distribution Detection

[41] Benchmark for Out-of-Distribution Detection in Deep Reinforcement Learning

[42] Semantically Coherent Out-of-Distribution Detection

[43] Fine-grain Inference on Out-of-Distribution Data with Hierarchical Classification

[44] Towards Few-shot Out-of-Distribution Detection

[45] Object Detectors in the Open Environment Challenges, Solutions, and Outlook

[46] General-Purpose Multi-Modal OOD Detection Framework

[47] Toward a Realistic Benchmark for Out-of-Distribution Detection

[48] How Good Are LLMs at Out-of-Distribution Detection

[49] DOI Divergence-based Out-of-Distribution Indicators via Deep Generative Models

[50] Out of Distribution Detection on ImageNet-O

[51] A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks

[52] Enhancing The Reliability of Out-of-distribution Image Detection in Neural Networks

[53] Explainable Deep One-Class Classification

[54] Gradient-based Novelty Detection Boosted by Self-supervised Binary Classification

[55] Mahalanobis-Aware Training for Out-of-Distribution Detection

[56] Learning a Neural-network-based Representation for Open Set Recognition

[57] Class Relevance Learning For Out-of-distribution Detection

[58] SSD A Unified Framework for Self-Supervised Outlier Detection

[59] ExCeL Combined Extreme and Collective Logit Information for Enhancing Out-of-Distribution Detection

[60] Towards Improved Variational Inference for Deep Bayesian Models

[61] Structured Output Learning with Conditional Generative Flows

[62] Generative Adversarial Networks

[63] A Simple Unified Framework for Anomaly Detection in Deep Reinforcement Learning

[64] Understanding the properties and limitations of contrastive learning for Out-of-Distribution detection

[65] Detecting Out-of-distribution Samples via Variational Auto-encoder with Reliable Uncertainty Estimation

[66] Outlier Exposure with Confidence Control for Out-of-Distribution Detection

[67] Shifting Transformation Learning for Out-of-Distribution Detection

[68] Boosting Out-of-distribution Detection with Typical Features

[69] Two-step counterfactual generation for OOD examples

[70] Using Semantic Information for Defining and Detecting OOD Inputs

[71] Large Class Separation is not what you need for Relational Reasoning-based OOD Detection

[72] Data Invariants to Understand Unsupervised Out-of-Distribution Detection

[73] OpenAUC Towards AUC-Oriented Open-Set Recognition

[74] In or Out Fixing ImageNet Out-of-Distribution Detection Evaluation

[75] Rethinking Out-of-distribution (OOD) Detection Masked Image Modeling is All You Need

[76] In Rain or Shine Understanding and Overcoming Dataset Bias for Improving Robustness Against Weather Corruptions for Autonomous Vehicles

[77] OOD-CV A Benchmark for Robustness to Out-of-Distribution Shifts of Individual Nuisances in Natural Images

[78] Taming False Positives in Out-of-Distribution Detection with Human Feedback

[79] NODI Out-Of-Distribution Detection with Noise from Diffusion

[80] AUTO Adaptive Outlier Optimization for Online Test-Time OOD Detection

[81] Improving Uncertainty-based Out-of-Distribution Detection for Medical Image Segmentation

[82] Rainproof An Umbrella To Shield Text Generators From Out-Of-Distribution Data

[83] Towards Rigorous Design of OoD Detectors

[84] Diffusion Denoised Smoothing for Certified and Adversarial Robust Out-Of-Distribution Detection

[85] Unsupervised Out-of-Distribution Detection with Batch Normalization

[86] Towards Open Set Deep Networks

[87] Mind the Gap

[88] Evaluating Out-of-Distribution Detectors Through Adversarial Generation of Outliers