International Journal of Computer Vision
https://doi.org/10.1007/s11263-024-02117-4

Generalized Out-of-Distribution Detection: A Survey

Jingkang Yang[1] · **Kaiyang Zhou**[1] · **Yixuan Li**[2] · **Ziwei Liu**[1]

Received: 27 April 2023 / Accepted: 26 April 2024

Abstract
Out-of-distribution (OOD) detection is critical to ensuring the reliability and safety of machine learning systems. For instance, in autonomous driving, we would like the driving system to issue an alert and hand over the control to humans when it detects unusual scenes or objects that it has never seen during training time and cannot make a safe decision. The term, OOD detection, first emerged in 2017 and since then has received increasing attention from the research community, leading to a plethora of methods developed, ranging from classification-based to density-based to distance-based ones. Meanwhile, several other problems, including anomaly detection (AD), novelty detection (ND), open set recognition (OSR), and outlier detection (OD), are closely related to OOD detection in terms of motivation and methodology. Despite common goals, these topics develop in isolation, and their subtle differences in definition and problem setting often confuse readers and practitioners. In this survey, we first present a unified framework called *generalized OOD detection*, which encompasses the five aforementioned problems, i.e.,AD, ND, OSR, OOD detection, and OD. Under our framework, these five problems can be seen as special cases or sub-tasks, and are easier to distinguish. Despite comprehensive surveys of related fields, the summarization of OOD detection methods remains incomplete and requires further advancement. This paper specifically addresses the gap in recent technical developments in the field of OOD detection. It also provides a comprehensive discussion of representative methods from other sub-tasks and how they relate to and inspire the development of OOD detection methods. The survey concludes by identifying open challenges and potential research directions.

Keywords Out-of-distribution detection · AI safety · Model trustworthiness · Open set recognition · Computer vision

1 Introduction

A trustworthy visual recognition system should not only produce accurate predictions on known context, but also detect unknown examples and reject them (or hand them over to human users for safe handling) (Amodei et al., 2016; Mohseni et al., 2021; Hendrycks et al., 2021; Hendrycks & Mazeika, 2022). For instance, a well-trained food classifier should be able to detect non-food images such as selfies uploaded by users, and reject such input instead of blindly classifying them into existing food categories. In safety-critical applications such as autonomous driving, the driving system must issue a warning and hand over the control to drivers when it detects unusual scenes or objects it has never seen during training (Fig. 1).

Most existing machine learning models are trained based on the closed-world assumption (Krizhevsky et al., 2012; He et al., 2015), where the test data is assumed to be drawn *i.i.d.* from the same distribution as the training data, known as in-distribution (ID). However, when models are deployed in an *open-world* scenario (Drummond & Shearer, 2006), test samples can be out-of-distribution (OOD) and therefore should be handled with caution. The distributional shifts can be caused by semantic shift (e.g.,OOD samples are drawn from different classes) (Hendrycks & Gimpel, 2017), or covariate shift (e.g.,OOD samples from a differ-

Communicated by Hong Liu.

✉ Ziwei Liu
ziwei.liu@ntu.edu.sg

Jingkang Yang
jingkang001@ntu.edu.sg

Kaiyang Zhou
kaiyang.zhou@ntu.edu.sg

Yixuan Li
sharonli@cs.wisc.edu

[1] S-Lab, Nanyang Technological University, Singapore, Singapore

[2] Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA

Published online: 23 June 2024

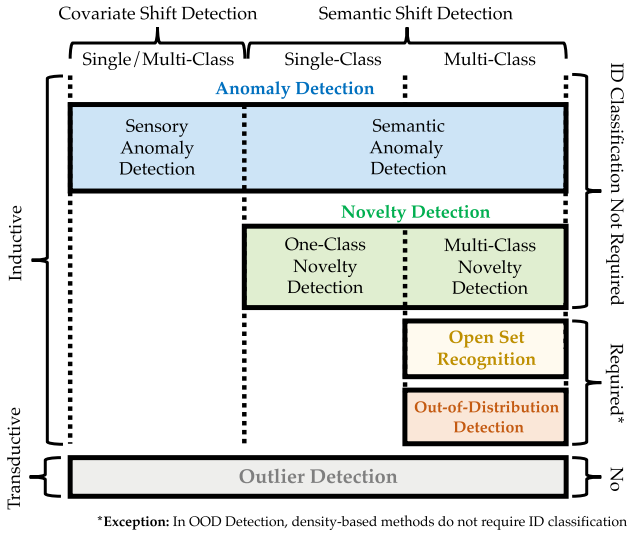

Fig. 1 Taxonomy of generalized OOD detection framework, illustrated by classification tasks. Four bases are used for the task taxonomy: (1) Distribution shift to detect: the task focuses on detecting covariate shift or semantic shift; (2) ID data type: the ID data contains one single class or multiple classes; (3) Whether the task requires ID classification; (4) Transductive learning task requires all observations; inductive tasks follow the train-test scheme. Note that ND is often interchangeable with AD, but ND is more concerned with semantic anomalies. OOD detection is generally interchangeable with OSR for classification tasks

ent domain) (Ben-David et al., 2010; Li et al., 2017b; Wang & Deng, 2018).

The detection of semantic distribution shift (e.g.,due to the occurrence of new classes) is the focal point of OOD detection tasks, where the label space $\mathcal{Y}$ can be different between ID and OOD data and hence the model should not make any prediction. In addition to OOD detection, several problems adopt the "open-world" assumption and have a similar goal of identifying OOD examples. These include outlier detection (OD) (Aggarwal & Yu, 2001; Hodge & Austin, 2004; Ben-Gal, 2005; Wang et al., 2019a), anomaly detection (AD) (Ruff et al., 2021; Pang et al., 2020; Bulusu et al., 2020; Chalapathy & Chawla, 2019), novelty detection (ND) (Pimentel et al., 2014; Miljković, 2010; Markou & Singh, 2003a; Markou & Singh, 2003b), and open set recognition (OSR) (Boult et al., 2019; Huang & Chen, 2020; Mahdavi & Carvalho, 2021). While all these problems are related to each other by sharing similar motivations, subtle differences exist among the *sub-topics* in terms of the specific definition. However, the lack of a comprehensive understanding of the relationship between the different sub-topics leads to confusion for both researchers and practitioners. Even worse, these sub-topics, which are supposed to be compared and learned from each other, are developing in isolation.

In this survey, we for the first time clarify the similarities and differences between these problems, and present a unified framework termed *generalized OOD detection*. Under this framework, the five problems (i.e.,AD, ND, OSR, OOD detection, and OD) can be viewed as special cases or sub-topics. While other sub-topics have been extensively surveyed, the summarization of OOD detection methods is still inadequate and requires further exploration. This paper fills this gap by focusing specifically on recent technical developments in OOD detection, analyzing fair experimental comparisons among classical methods on common benchmarks. Our survey concludes by highlighting open challenges and outlining potential avenues for future research.

We further conduct a literature review for each sub-topic, with a special focus on the OOD detection task. To sum up, we make three contributions to the research community:

1. **A Unified Framework** For the first time, we systematically review five closely related topics of AD, ND, OSR, OOD detection, and OD, and present a unified framework of *generalized OOD detection*. Under this framework, the similarities and differences of the five sub-topics can be systematically compared and analyzed. We hope our unification helps the community better understand these problems and correctly position their research in the literature.
2. **A Comprehensive Survey for OOD Detection** Noticing the existence of comprehensive surveys on AD, ND, OSR, and OD methodologies in recent years (Ruff et al., 2021; Pang et al., 2020; Bulusu et al., 2020; Chalapathy & Chawla, 2019; Huang & Chen, 2020), this survey provides a comprehensive overview of OOD detection methods and thus complements existing surveys. By connecting with methodologies of other sub-topics that are also briefly reviewed, as well as sharing the insights from a fair comparison on a standard benchmark, we hope to provide readers with a more holistic understanding of the developments for each problem and their interconnections, especially for OOD detection.
3. **Future Research Directions** We draw readers' attention to some problems or limitations that remain in the current generalized OOD detection field. We conclude this survey with discussions on open challenges and opportunities for future research.

Organization of Remaining Sections To match the mentioned contributions, in Sect. 2, we introduce the generalized OOD detection framework and discuss related topics to better position our survey. In Sect. 3, we provide a comprehensive overview of the methodologies for OOD detection, categorizing them into four groups: (1) classification-based methods, which largely rely on classifiers; (2) density-based methods, which detect OOD by modeling data density; (3) distance-based methods, which use distance metrics (usually in the feature space) to identify OODs; and (4) reconstruction-

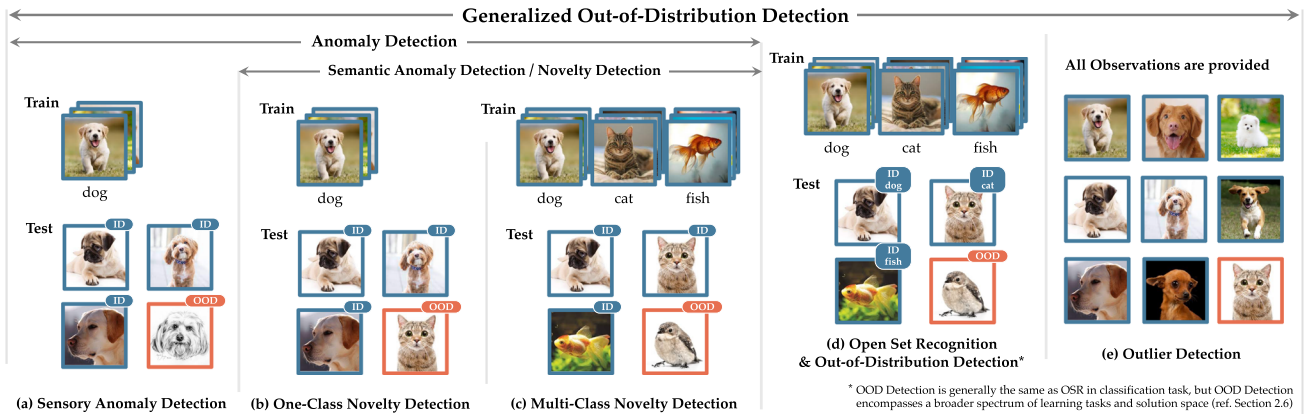

Fig. 2 Illustration of sub-tasks under generalized OOD detection framework with vision tasks. Tags on test images refer to model's expected predictions. **a** In *sensory anomaly detection*, test images with covariate shift will be considered as OOD. No semantic shift occurs in this setting. **b** In *one-class novelty detection*, normal/ID images belong to one class. Test images with semantic shift will be considered as OOD. **c** In *multi-class novelty detection*, ID images belong to multiple classes. Test images with semantic shift will be considered as OOD. Note that **b** and **c** compose novelty detection, which is identical to the topic of semantic anomaly detection. **d** *Open set recognition* is identical to multi-class novelty detection in the task of detection, with the only difference that open set recognition further requires ID classification. *Out-of-distribution detection* solves the same problem as open-set recognition. It canonically aims to detect test samples with semantic shift without losing the ID classification accuracy. However, OOD Detection encompasses a broader spectrum of learning tasks and solution space. **e** *Outlier detection* does not follow a train-test scheme. All observations are provided. It fits in the generalized OOD detection framework by defining the majority distribution as ID. Outliers can have any distribution shift from the majority

based methods, which are characterized by reconstruction techniques. In Sect. 4, we briefly introduce methodologies for other sub-tasks, including AD, ND, OSR, and OD, to provide readers with a broader understanding of OOD-related problems and inspire the development of more effective methods. To offer readers further insights from an empirical perspective, in Sect. 5 we conduct a thorough analysis that provides a fair comparison between representative OOD detection methods and methods from other sub-tasks. Additionally, we highlight some of the remaining problems and limitations that exist in the current generalized OOD detection field. We conclude this survey with a discussion on the open challenges and opportunities for future research. It is worth noting that a concurrent survey (Salehi et al., 2021) provides a detailed explanation of OOD-related methods, which greatly complements our work.

2 Generalized OOD Detection

Framework Overview In this section, we introduce a unified framework termed *generalized OOD detection*, which encapsulates five related sub-topics: anomaly detection (AD), novelty detection (ND), open set recognition (OSR), out-of-distribution (OOD) detection, and outlier detection (OD). These sub-topics can be similar in the sense that they all define a certain *in-distribution*, with the common goal of detecting *out-of-distribution* samples under the open-world assumption. However, subtle differences exist among the sub-topics in terms of the specific definition and properties of ID and OOD data—which are often overlooked by the research community. To this end, we provide a clear introduction and description of each sub-topic in respective subsections (from Sect. 2.1 to 2.5). Each subsection details the motivation, background, formal definition, as well as relative position within the unified framework. Applications and benchmarks are also introduced, with concrete examples that facilitate understanding. Figure 2 illustrates the settings for each sub-topic. In the end, we conclude this section by introducing the neighborhood topics to clarify the scope of the generalized OOD detection framework (Sect. 2.6).

Preliminary: Distribution Shift In our framework, we recognize the complexity and interconnectedness of distribution shifts, which are central to understanding various OOD scenarios. Distribution shifts can be broadly categorized into *covariate shift* and *semantic (label) shift*, but it's important to clarify their interdependence. Firstly, let's define the input space as $\mathcal{X}$ (sensory observations) and the label space as $\mathcal{Y}$ (semantic categories). The data distribution is represented by the joint distribution $P(X, Y)$ over the space $\mathcal{X} \times \mathcal{Y}$. Distribution shift can occur in either the marginal distribution $P(X)$, or both $P(Y)$ and $P(X)$. Note that shift in $P(Y)$ naturally triggers shift in $P(X)$.

Covariate Shift This occurs when there is a change in the marginal distribution $P(X)$, affecting the input space, while the label space $\mathcal{Y}$ remains constant. Examples of covariate distribution shift on $P(X)$ include adversarial examples (Goodfellow et al., 2015; Madry et al., 2018), domain shift (Qui nonero-Candela et al., 2009), and style changes (Gatys et al., 2016).

Semantic Shift This involves changes in both $P(Y)$ and indirectly $P(X)$. A shift in the label space $P(Y)$ implies the introduction of new categories or the alteration of existing ones. This change naturally affects the input space $P(X)$ since the nature of the data being observed or collected is now different.

Remark Given the interdependence between $P(X)$ and $P(Y)$, it's crucial to distinguish the intentions behind different types of distribution shifts. We define *Covariate Shift* as scenarios where changes are intended in the input space ($P(X)$) without any deliberate alteration to the label space ($P(Y)$). On the other hand, *Semantic Shift* specifically aims to modify the semantic content, directly impacting the label space ($P(Y)$) and, consequently, the input space ($P(X)$).

Importantly, we note that covariate shift is more commonly used to evaluate model *generalization* and robustness performance, where the label space $\mathcal{Y}$ remains the same during test time. On the other hand, the detection of semantic distribution shift (e.g.,due to the occurrence of new classes) is the focal point of many *detection* tasks considered in this framework, where the label space $\mathcal{Y}$ can be different between ID and OOD data and hence the model should not make any prediction.

With the concept of distribution shift in mind, readers can get a general idea of the differences and connections among sub-topics/tasks in Fig. 1. Notice that different sub-tasks can be easily identified with the following four dichotomies: (1) covariate/semantic shift dichotomy; (2) single/multiple class dichotomy; (3) ID classification needed/non-needed dichotomy; (4) inductive/transductive dichotomy. Next, we proceed with elaborating on each sub-topic.

2.1 Anomaly Detection

Background The notion of "anomaly" stands in contrast with the "normal" defined in advance. The concept of "normal" should be clear and reflect the real task. For example, to create a "not-hotdog detector", the concept of the normal should be clearly defined as the hotdog class, i.e.,a food category, so that objects that violate this definition are identified as anomalies, which include steaks, rice, and non-food objects like cats and dogs. Ideally, "hotdog" would be regarded as a homogeneous concept, regardless of the sub-classes of French or American hotdog.

Current anomaly detection settings often restrict the environment of interest to some specific scenarios. For example, the "not-hotdog detector" only focuses on realistic images, assuming the nonexistence of images from other domains such as sketches. Another realistic example is industrial defect detection, which is based on only one set of assembly lines for a specific product. In other words, the "open-world" assumption is usually not completely "open". Nevertheless, "not-hotdog" or "defects" can form a large unknown space that breaks the "closed-world" assumption.

In summary, the key to anomaly detection is to define normal clearly (usually without sub-classes) and detect all possible anomalous samples under some specific scenarios.

Definition Anomaly detection (AD) (Chandola et al., 2009) aims to detect any anomalous samples that deviate from the predefined normality during testing. The deviation can happen due to either covariate shift or semantic shift, which leads to two sub-tasks: sensory AD and semantic AD, respectively (Ruff et al., 2021).

Sensory AD detects test samples with covariate shift, under the assumption that normalities come from the same covariate distribution. No semantic shift takes place in sensory AD settings. On the other hand, semantic AD detects test samples with label shift, assuming that normalities come from the same semantic distribution (category), i.e.,normalities should belong to only one class.

Formally, in sensory AD, normalities are from in-distribution $P(X)$ while anomalies encountered at test time are from out-of-distribution $P'(X)$, where $P(X) \neq P'(X)$—only covariate shift occurs. The goal in sensory AD is to detect samples from $P'(X)$. No semantic shift occurs in this setting, i.e.,$P(Y) = P'(Y)$. Conversely, for semantic AD, only semantic shift occurs (i.e.,$P(Y) \neq P'(Y)$) and the goal is to detect samples that belong to novel classes.

Remark: Sensory/Semantic Dichotomy Our sensory/semantic dichotomy for the AD sub-task definition comes from the low-level sensory anomalies and high-level semantic anomalies that are introduced in Ahmed and Courville (2020) and highlighted in the recent AD survey (Ruff et al., 2021), to reflect the rise of deep learning. Note that although most sensory and semantic AD methods are shown to be mutually inclusive due to the common shift on $P(X)$, some approaches are specialized in one of the sub-tasks (ref. Sect. 4.2). Recent research communities are also trending on subdividing types of anomalies to develop targeted methods, so that practitioners can select the optimal solution for their own practical problem (Ahmed & Courville, 2020; Zhang et al., 2021).

Position in Framework Under the generalized OOD detection framework, the definition of "normality" seamlessly connects to the notion of "in-distribution", and "anomaly" corresponds to "out-of-distribution". Importantly, AD treats ID samples as a whole, which means that regardless of the number of classes (or statistical modalities) in ID data, AD does not require differentiation in the ID samples. This feature is an important distinction between AD and other sub-topics such as OSR and OOD detection.

Application and Benchmark Sensory AD only focuses on objects with the same or similar semantics, and identifies the observational differences on their surface. Samples with sensory differences are recognized as sensory anomalies. Example applications include adversarial defense (Akhtar

& Mian, 2018), forgery recognition of biometrics and artworks (Patel et al., 2016; Wen et al., 2015; Nixon et al., 2008; Polatkan et al., 2009), image forensics (Dolhansky et al., 2019; Jiang et al., 2021c; Yang et al., 2020c), industrial inspection (Bergmann et al., 2019; Chu & Kitani, 2020; Atha & Jahanshahi, 2018), etc.. The most popular academic AD benchmark is MVTec-AD (Bergmann et al., 2019) for industrial inspection. Beyond academic research, sensory AD has significant potential in various real-world applications, such as detecting counterfeit items in retail and e-commerce and identifying manipulated media in journalism and law enforcement.

In contrast to sensory AD, semantic AD only focuses on the semantic shift. An example of real-world applications is crime surveillance (Idrees et al., 2018; Diehl & Hampshire, 2002). Active image crawlers for a specific category also need semantic AD methods to ensure the purity of the collected images (Li & Fei-Fei, 2010). An example of the academic benchmarks is to recursively use one class from MNIST as ID during training, and ask the model to distinguish it from the rest of the 9 classes during testing.

Evaluation In the AD benchmarks, test samples are annotated to be either normal or abnormal. The deployed anomaly detector will produce a confidence score for a test sample, indicating how confident the model considers the sample as normality. Samples below the predefined confidence threshold are considered abnormal. By viewing the anomalies as positive and true normalities as negative,[1] different thresholds will produce a series of true positive rates (TPR) and false-positive rates (FPR)—from which we can calculate the area under the receiver operating characteristic curve (AUROC) (Fawcett, 2006). Similarly, the precision and recall values can be used to compute metrics of F-scores and the area under the precision-recall curve (AUPR) (Powers, 2020). Note that there can be two variants of AUPR values: one treating "normal" as the positive class, and the other treating "abnormal" as the positive class. For AUROC and AUPR, a higher value indicates better detection performance.

Remark: Alternative Taxonomy on Anomalies Some previous literature considers anomalies types to be three-fold: point anomalies, conditional or contextural anomalies, and group or collective anomalies (Pang et al., 2020; Chalapathy & Chawla, 2019; Ruff et al., 2021). In this survey, we mainly focus on point anomalies detection for its popularity in practical applications and its adequacy to elucidate the similarities and differences between sub-tasks. Details of the other two kinds of anomalies, i.e.,contextural anomalies that often occur in time-series tasks, and collective anomalies that are common in the data mining field, are not covered in this survey. We recommend readers to the recent AD survey papers (Ruff et al., 2021) for an in-depth discussion on them.

Remark: Taxonomy Based on Supervision We use sensory/semantic dichotomy to subdivide AD at the task level. From the perspective of methodologies, some literature categorizes AD techniques into unsupervised and (semi-) supervised settings. Note that these two taxonomies are orthogonal as they focus on tasks and methods respectively.

2.2 Novelty Detection

Background The word "novel" generally refers to the unknown, new, and something interesting. While novelty detection (ND) is often interchangeable with AD in the community, strictly speaking, their subtle difference is worth noticing. In terms of motivation, novelty detection usually does not perceive "novel" test samples as erroneous, fraudulent, or malicious as AD does, but cherishes them as learning resources for potential future use with a positive learning attitude (Pang et al., 2020; Ruff et al., 2021). In fact, novelty detection is also known as "novel class detection" (Markou & Singh, 2003a, b), indicating that it is primarily focusing on detecting semantic shift.

Definition Novelty detection aims to detect any test samples that do not fall into any training category. The detected novel samples are usually prepared for future constructive procedures, such as more specialized analysis, or incremental learning of the model itself. Based on the number of training classes, ND contains two different settings: (1) one-class novelty detection (*one-class ND*): only one class exists in the training set; (2) multi-class novelty detection (*multi-class ND*): multiple classes exist in the training set. It is worth noting that despite having many ID classes, the goal of multi-class ND is only to distinguish novel samples from ID. Both one-class and multi-class ND are formulated as binary classification problems.

Position in Framework Under the generalized OOD detection framework, ND deals with the setting where OOD samples have semantic shift, without the need for classification in the ID set even if possible. Therefore, ND shares the same problem definition with semantic AD.

Application and Benchmark Real-world ND application includes video surveillance (Idrees et al., 2018; Diehl & Hampshire, 2002), planetary exploration (Kerner et al., 2019) and incremental learning (Al-Behadili et al., 2015; Pathak et al., 2017). For one-class ND, an example academic benchmark can be identical to that of semantic AD, which considers one class from MNIST as ID and the rest as the novel. The corresponding MNIST benchmark for multi-class ND may use the first 6 classes during training, and test on the remaining 4 classes as OOD. Beyond academic research, ND has significant potential on new drag discovery, new species discovery, etc.

[1] Align with MSP (Hendrycks & Gimpel, 2017) Check https://github.com/Jingkang50/OpenOOD/issues/206this issue in OpenOOD.

Evaluation The evaluation of ND is identical to AD, which is based on AUROC, AUPR, or F-scores (see details in Sect. 2.1).

Remark: One-Class/Multi-class Dichotomy Although the ND models do not require the ID classification even with multi-class annotations, the method on multi-class ND can be different from one-class ND, as multi-class ND can make use of the multi-class classifier while one-class ND cannot. Also note that semantic AD can be further split into one-class semantic AD and multi-class semantic AD that matches ND, as semantic AD is equivalent to ND.

Remark: Nuance Between AD and ND Apart from the special interest in semantics, some literature (Perera et al., 2019; Xia et al., 2015) also point out that ND is supposed to be fully unsupervised (no novel data in training), while AD might have some abnormal training samples. It's important to note that neither AD nor ND necessitates the classification of ID data. This is a key distinction between OSR and OOD detection, which we will discuss in subsequent sections.

2.3 Open Set Recognition

Background Machine learning models trained in the closed-world setting can incorrectly classify test samples from unknown classes as one of the known categories with high confidence (Scheirer et al., 2013). Some literature refers to this notorious overconfident behavior of the model as "arrogance", or "agnostophobia" (Dhamija et al., 2018). Open set recognition (OSR) is proposed to address this problem, with their own terminology of "known known classes" to represent the categories that exist at training, and "unknown unknown classes" for test categories that do not fall into any training category. Some other terms, such as open category detection (Liu et al., 2018a) and open set learning (Fang et al., 2021), are simply different expressions for OSR.

Definition Open set recognition requires the multi-class classifier to simultaneously: (1) accurately classify test samples from "known known classes", and (2) detect test samples from "unknown unknown classes".

Position in Framework OSR well aligns with our generalized OOD detection framework, where "known known classes" and "unknown unknown classes" correspond to ID and OOD respectively. Formally, OSR deals with the case where OOD samples during testing have semantic shift, i.e.,$P(Y) \neq P'(Y)$. The goal of OSR is largely shared with that of multi-class ND—the only difference is that OSR additionally requires accurate classification of ID samples from $P(Y)$.

Application and Benchmark OSR supports the robust deployment of real-world image classifiers in general, which can reject unknown samples in the open world (Sorio et al., 2010; Xu et al., 2019). An example academic benchmark on MNIST can be identical to multi-class ND, which considers the first 6 classes as ID and the remaining 4 classes as OOD. In addition, OSR further requires a good classifier on the 6 ID classes.

Evaluation Similar to AD and ND, the metrics for OSR include F-scores, AUROC, and AUPR. Beyond them, the classification performance is also evaluated by standard ID accuracy. While the above metrics evaluate the novelty detection and ID classification capabilities independently, some works raise some evaluation criteria for joint evaluation, such as CCR@FPRx (Dhamija et al., 2018), which calculates the class-wise recall when a certain FPR equal to x (e.g.,10^{-1}) is achieved.

2.4 Out-of-Distribution Detection

Background With the observation that deep learning models are often inappropriate but in fact overconfident in classifying samples from different semantic distributions in the image classification task and text categorization (Hendrycks & Gimpel, 2017), the field of out-of-distribution detection emerges, requiring the model to reject inputs that are semantically different from the training distribution and therefore should not be predicted by the model.

Definition Out-of-distribution detection, or OOD detection, aims to detect test samples drawn from a distribution that is different from the training distribution, with the definition of distribution to be well-defined according to the application in the target. For most machine learning tasks, the distribution should refer to "label distribution", which means that OOD samples should not have overlapping labels w.r.t.training data. Formally, in the OOD detection, the test samples come from a distribution whose semantics are shifted from ID, i.e.,$P(Y) \neq P'(Y)$. Note that the training set usually contains multiple classes, and OOD detection should NOT harm the ID classification capability.

Position in Framework Out-of-distribution detection can be canonical to OSR in common machine learning tasks like multi-class classification—keeping the classification performance on test samples from ID class space $\mathcal{Y}$, and reject OOD test samples with semantics outside the support of $\mathcal{Y}$. Also, the multi-class setting and the requirement of ID classification distinguish the task from AD and ND.

Application and Benchmark The application of OOD detection usually falls into safety-critical situations, such as autonomous driving (Huang et al., 2020b; Geiger et al., 2012). In the context of self-driving vehicles, OOD detection (also OSR) plays a crucial role in identifying novel or unexpected objects and scenarios, allowing the system to take appropriate actions to ensure the safety of passengers and pedestrians. Other potential real-world applications include medical diagnosis, where OOD detection can help flag unusual patient cases that may require further attention from healthcare professionals, and industrial monitoring,

where it can identify anomalies in sensor data that could indicate potential equipment failures or safety hazards. An example academic benchmark is to use CIFAR-10 as ID during training and to distinguish CIFAR images from other datasets such as SVHN, etc.. Researchers should pay attention that OOD datasets should NOT have label overlapping with ID datasets when building the benchmark.

Evaluation Apart from F-scores, AUROC, and AUPR, another commonly-used metric is FPR@TPRx, which measures the FPR when the TPR is x (e.g.,0.95). Some works also use an alternative metric, TNR@TPRx, which is equivalent to 1-FPR@TPRx. OOD detection also concerns the performance of ID classification.

Remark: OSR vs **OOD Detection** The difference between OSR and OOD detection tasks is three-fold.

(1) Different Benchmark Setup OSR benchmarks usually split one multi-class classification dataset into ID and OOD parts according to classes, while OOD detection takes one dataset as ID and finds several other datasets as OOD with the guarantee of non-overlapping categories between ID/OOD datasets. However, despite the different benchmark traditions of the two sub-tasks, they are in fact tackling the same problem of semantic shift detection.

(2) No Additional data in OSR Due to the requirement of theoretical open-risk bound guarantee, OSR discourages the usage of additional data during training by design (Boult et al., 2019). This restriction precludes methods that are more focused on effective performance improvements (e.g.,outlier exposures (Hendrycks et al., 2019b; Zhang et al., 2023b)) but may violate OSR constraints.

(3) Broadness of OOD Detection Compare to OSR, OOD detection encompasses a broader spectrum of learning tasks (e.g.,multi-label classification (Hendrycks et al., 2022a)), wider solution space (to be discussed in Sect. 3).

Remark: Mainstream OOD Detection Focuses on Semantics While most works in the current community interpret the keyword "out-of-distribution" as "out-of-label/semantic-distribution", some OOD detection works also consider detecting covariate shifts (Hsu et al., 2020), which claim that covariate shift usually leads to a significant drop in model performance and therefore needs to be identified and rejected. However, although detecting covariate shift is reasonable on some specific tasks (usually due to high-risk or privacy reasons) that are to be discussed in the following paragraph, research on this topic remains a controversial task *w.r.t* OOD generalization tasks (c.f.Sects. 2.6 and 6.2). Detecting semantic shift has been the mainstream of OOD detection tasks.

Remark: To Generalize, or To Detect? We provide another definition from the perspective of generalization: Out-of-distribution detection, or OOD detection, aims to detect test samples to which the model cannot or does not want to generalize (Pleiss et al., 2019). In most of the machine learning tasks, such as image classification, the models are expected to generalize their prediction capability to samples with covariate shift, and they are only unable to generalize when semantic shift occurs. However, for applications where models are by-design nontransferable to other domain, such as many deep reinforcement learning tasks like game AI (Vinyals et al., 2017; Sedlmeier et al., 2019), the key term "distribution" should refer to "data/input distribution", so that the model should refuse to decide the environment that is not the same as the training environment, i.e.,$P(X) \neq P'(X)$. Similar applications are those high-risk tasks such as medical image classification (Zimmerer et al., 2022) or in privacy-sensitive scenario (Tariq et al., 2020), where the models are expected to be very conservative and only make predictions for samples exactly from the training distribution, rejecting any samples that deviate from it. Recent studies (Averly & Chao, 2023) also highlight a model-specific view: a robust model should generalize to examples with covariate shift; a weak model should reject them. Ultimately, an OOD detection task is considered valid when it successfully balances the aspects of "detection" and "generalization", taking into account factors such as meaningfulness and the inherent challenges presented by the task. Nonetheless, detecting semantic shift remains the primary focus of OOD detection tasks and is central to this survey.

2.5 Outlier Detection

Background According to *Wikipedia* (Wikipedia contributors, 2021), an outlier is a data point that differs significantly from other observations. Recall that the problem settings in AD, ND, OSR, and OOD detect unseen test samples that are different from the training data distribution. In contrast, outlier detection directly processes all observations and aims to select outliers from the contaminated dataset (Ben-Gal, 2005; Hodge & Austin, 2004; Aggarwal & Yu, 2001). Since outlier detection does not follow the train-test procedure but has access to all observations, approaches to this problem are usually transductive rather than inductive (Bianchini et al., 2016).

Definition Outlier detection aims to detect samples that are markedly different from the others in the given observation set, due to either covariate or semantic shift.

Position in Framework Different from all previous subtasks, whose in-distribution is defined during training, the "in-distribution" for outlier detection refers to the majority of the observations. Outliers may exist due to semantic shift on $P(Y)$, or covariate shift on $P(X)$.

Application and Benchmark While mostly applied in data mining tasks (Ben-Gal, 2005; Basu & Meckesheimer, 2007; Dou et al., 2019), outlier detection is also used in real-world computer vision applications such as video surveillance (Xiao et al., 2015) and dataset cleaning (Liu et al.,

2004; Loureiro et al., 2004; Van den Broeck et al., 2005). For the application of dataset cleaning, outlier detection is usually used as a pre-processing step for the main tasks such as learning from open-set noisy labels (Wang et al., 2018), webly supervised learning (Chen & Gupta, 2015), and open-set semi-supervised learning (Cao et al., 2021). To construct an outlier detection benchmark on MNIST, one class should be chosen so that all samples that belong to this class are considered as inliers. A small fraction of samples from other classes are introduced as outliers to be detected.

Evaluation Apart from F-scores, AUROC, and AUPR, the evaluation of outlier detectors can be also evaluated by the performance of the main task it supports. For example, if an outlier detector is used to purify a dataset with noisy labels, the performance of a classifier that is trained on the cleaned dataset can indicate the quality of the outlier detector.

Remark: On Inclusion of Outlier Detection Interestingly, the outlier detection task can be considered as an outlier in the generalized OOD detection framework, since outlier detectors are operated on the scenario when all observations are given, rather than following the training-test scheme. Also, publications exactly on this topic are rarely seen in the recent deep learning venues. However, we still include outlier detection in our framework, because intuitively speaking, outliers also belong to one type of out-of-distribution, and introducing it can help familiarize readers more with various terms (e.g.,OD, AD, ND, OOD) that have confused the community for a long while.

2.6 Related Topics

Apart from the five sub-topics that are described in our *generalized OOD detection* framework (shown in Fig. 1), we further briefly discuss five related topics below, which help clarify the scope of this survey.

Learning with Rejection (LWR) LWR (Bartlett & Wegkamp, 2008) can date back to early works on abstention (Chow, 1970; Fumera & Roli, 2002), which considered simple model families such as SVMs (Cortes & Vapnik, 1995). The phenomenon of neural networks' overconfidence in OOD data is first revealed by Nguyen et al. (2015). Despite methodologies differences, subsequent works developed on OOD detection and OSR share the underlying spirit of classification with the rejection option.

Domain Adaptation/Generalization Domain Adaptation (DA) (Wang & Deng, 2018) and Domain Generalization (DG) (Zhou et al., 2021b) also follow "open-world" assumption. Different from generalized OOD detection settings, DA/DG expects the existence of covariate shift during testing without any semantic shift and requires classifiers to make accurate predictions into the same set of classes (Liu et al., 2020c). Noticing that OOD detection commonly concerns detecting the semantic shift, which is complementary to DA/DG. In the case when both covariate and semantic shift take place, the model should be able to detect semantic shift while being robust to covariate shift. More discussion on relations between DA/DG and OOD detection is in Sect. 6.2. The difference between DA and DG is that while the former requires extra but few training samples from the target domain, the latter does not.

Novelty Discovery Novelty discovery (Han et al., 2019; Zhao & Han, 2021; Jia et al., 2021; Vaze et al., 2022a; Joseph et al., 2022) requires all observations to be given in advance as outlier detection does. The observations are provided in a semi-supervised manner, and the goal is to explore and discover the new categories and classes in the unlabeled set. Different from outlier detection where outliers are sparse, the unlabeled set in novelty discovery setting can mostly consist of, and even be overwhelmed by unknown classes.

Zero-Shot Learning Zero-shot learning (Wang et al., 2019b) has a similar goal of novelty discovery but follows the training–testing scheme. The test set is under the "open-world" assumption with unknown classes, which expects classifiers trained only on the known classes to perform classification on unknown testing samples with the help of extra information such as label relationships.

Open-World Recognition Open-world recognition (Bendale & Boult, 2015) aims to build a lifelong learning machine that can actively detect novel images (Liu et al., 2019), label them as new classes, and perform continuous learning. It can be viewed as a combination of novelty detection (or open-set recognition) and incremental learning. More specifically, open-world recognition extends the concept of OSR by adding the ability to incrementally learn new classes over time. In open-world scenarios, the system not only identifies unknown instances but also can update its model to include these new classes as part of the known set. This approach is more dynamic and suited for real-world applications where the environment is not static, and new categories can emerge after the initial training phase (Parmar et al., 2023).

Conformal Prediction Conformal prediction (CP) stands as a robust statistical framework in machine learning, primarily designed to provide confidence measures for predictions (Shafer & Vovk, 2008; Angelopoulos & Bates, 2021). Distinctively, it yields prediction intervals with specified confidence levels, transcending the limitations of mere point estimates. In scenarios of OOD detection, the conformal prediction framework becomes particularly insightful: wider prediction intervals or lower confidence levels generated by conformal prediction methods can serve as indicators of such OOD data. Although research at the intersection of CP and OOD detection is still emerging (Kaur et al., 2022a, b; Cai et al., 2021), the potential of applying the conformal prediction framework in this domain is significant and warrants further exploration (Table 1; Fig. 3).

Table 1 Paper list for out-of-distribution detection

Sections	References
Section 3.1 Classification	
Section 3.1.1 Output-based Methods	
a: Training-free	Hendrycks and Gimpel (2017), Liang et al. (2018), Lee et al. (2018b), Liu et al. (2020), Sastry and Oore (2020), Wang et al. (2021), Zhang et al. (2023a), Sun et al. (2021b), Dong et al. (2022), Sun and Li (2022), Sun et al. (2022), Lin et al. (2021), Sastry and Oore (2019), Zhang et al. (2023a), Djurisic et al. (2023), Park et al. (2023b), Park et al. (2023a), Jiang et al. (2023b), Liu et al. (2023)
b: Training-based	DeVries and Taylor (2018), Wang et al. (2021, 2022b), Vyas et al. (2018), Bitterwolf et al. (2020), Chen et al. (2020b), Hein et al. (2019), Choi and Chung (2020), Chen et al. (2021c), Hein et al. (2019), Thulasidasan et al. (2019), Yun et al. (2019), DeVries and Taylor (2017), Hendrycks et al. (2019c, 2022c), Tack et al. (2020) , Meinke and Hein (2019), Bibas et al. (2021), Lin et al. (2021), Dong et al. (2022), Hsu et al. (2020), Wei et al. (2022), Lee et al. (2018c), Huang and Li (2021), Linderman et al. (2023), Shalev et al. (2018), Fort et al. (2021), Gan (2021)
Section 3.1.2 Outlier Exposure	
a: Real Outliers	Hendrycks et al. (2019b), Dhamija et al. (2018), Yu and Aizawa (2019), Mohseni et al. (2020), Chen et al. (2021c), Thulasidasan et al. (2021), Papadopoulos et al. (2021), Chen et al. (2021c), Ming et al. (2022b), Li and Vasconcelos (2020), Zhang et al. (2023b), Yang et al. (2021), Lu et al. (2023), Shafaei et al. (2019), Katz-Samuels et al. (2022), Wang et al. (2023b)
b: Data Generation	Lee et al. (2018a), Vernekar et al. (2019), Sricharan et al. (2018), Jeong and Kim (2020), Du et al. (2022b), Tao et al. (2023), Wang et al. (2023c), Zheng et al. (2023), Du et al. (2022a)
Section 3.1.3: Gradient-based Methods	Liang et al. (2018), Huang et al. (2021), Igoe et al. (2022)
Section 3.1.4: Bayesian Models	Gal and Ghahramani (2016), Lakshminarayanan et al. (2017), Osawa et al. (2019), Malinin and Gales (2018, 2019), Nandy et al. (2020), Kim et al. (2021)
Section 3.1.5: OOD for Foundation Models	Hendrycks et al. (2019a, 2020), Fort et al. (2021), Ming and Li (2023), Miyai et al. (2023a, b), Lu et al. (2023), Esmaeilpour et al. (2022), Ming et al. (2022a), Wang et al. (2023a)
Section 3.2: Density-based Methods	Zong et al. (2018), Abati et al. (2019), Pidhorskyi et al. (2018), Deecke et al. (2018), Sabokrou et al. (2018), Lee et al. (2018b), Kobyzev et al. (2020), Zisselman and Tamar (2020), Kingma and Dhariwal (2018), Van Oord et al. (2016), Jiang et al. (2021a), Nalisnick et al. (2018), Choi et al. (2018), Kirichenko et al. (2020), Ren et al. (2019), Serrà et al. (2020), Xiao et al. (2020), Wang et al. (2022a)
Section 3.3: Distance-based Methods	Lee et al. (2018b), Ren et al. (2021), Techapanurak et al. (2020), Chen et al. (2020c), Zaeemzadeh et al. (2021), Van Amersfoort et al. (2020), Huang et al. (2020a), Sun et al. (2022), Ming et al. (2023), Kim et al. (2023)
Section 3.4: Reconstruction-based Methods	Denouden et al. (2018), Zhou (2022), Yang et al. (2022c), Jiang et al. (2023a), Li et al. (2023b)
Section 3.5: Theoretical Analysis	Zhang et al. (2021), Morteza and Li (2022), Scheirer et al. (2013), Jain et al. (2014), Rudd et al. (2017), Liu et al. (2018a), Fang et al. (2021, 2022)

3 OOD Detection: Methodology

In this section, we introduce the methodology for OOD detection. Initially, we explore classification-based models in Sect. 3.1. These models primarily utilize the model's output, such as softmax scores, to identify OOD instances. We further examine outlier exposure-based methods that leverage external data sources and other types of methods. The later section is followed by density-based methods in Sect. 3.2. Distance-based methods will be introduced in Sect. 3.3. A brief discussion will be included at the end.

3.1 Classification-Based Methods

Research on OOD detection originated from a simple baseline, that is, using the maximum softmax probability as the indicator score of ID-ness (Hendrycks & Gimpel, 2017). Early OOD detection methods focus on deriving improved OOD scores based on the output of neural networks.

3.1.1 Output-Based Methods

a. Post-Hoc Detection Post-hoc methods have the advantage of being easy to use without modifying the training

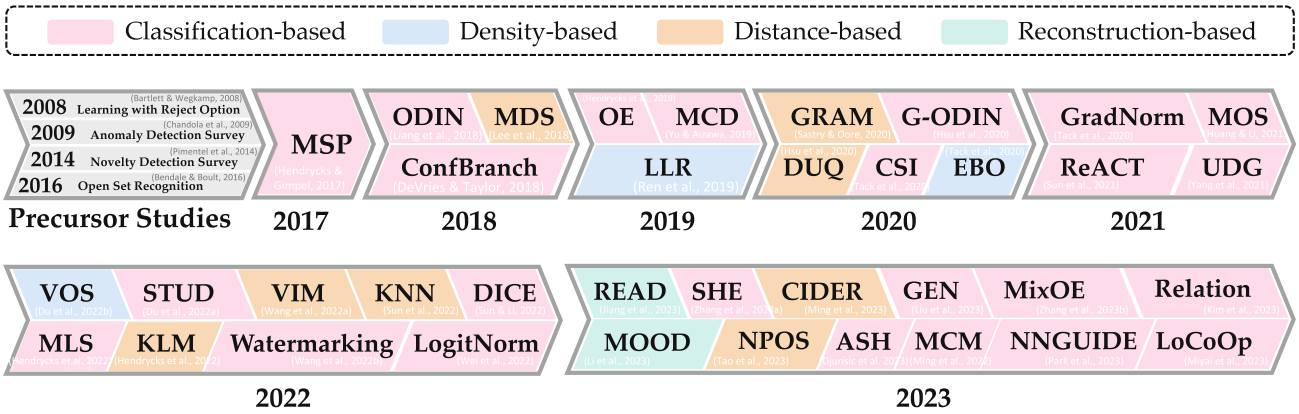

Fig. 3 Timeline for representative OOD detection methodologies. Different colors indicate different categories of methodologies. Each method has its corresponding reference (inconspicuous white) in the lower right corner. Methods with high citations and open-source code are prioritized for inclusion in this figure (Color figure online)

procedure and objective. The property can be important for the adoption of OOD detection methods in real-world production environments, where the overhead cost of retraining can be prohibitive. Early work ODIN (Liang et al., 2018) is a post-hoc method that uses temperature scaling and input perturbation to amplify the ID/OOD separability. Key to the method, a sufficiently large temperature has a strong smoothing effect that transforms the softmax score back to the logit space—which effectively distinguishes ID vs. OOD. Note that this is different from confidence calibration, where a much milder T is employed. While calibration focuses on representing the true correctness likelihood of ID data only, the ODIN score is designed to maximize the gap between ID and OOD data and may no longer be meaningful from a predictive confidence standpoint. Built on the insights, recent work (Liu et al., 2020; Lin et al., 2021) proposed using an energy score for OOD detection, which enjoys theoretical interpretation from a likelihood perspective (Morteza & Li, 2022). Test samples with lower energy are considered ID and vice versa. JointEnergy score (Wang et al., 2021) is then proposed to perform OOD detection for multi-label classification networks. The most recent work SHE (Zhang et al., 2023a) uses stored patterns that represent classes to measure the discrepancy of unseen data for OOD detection, which is hyperparameter-free and computationally efficient compared to classic energy methods. Techniques such as layer-wise Mahalanobis distance (Lee et al., 2018b) and Gram Matrix (Sastry & Oore, 2020) are implemented for better-hidden feature quality to perform density estimation.

Recently, one fundamental cause of the overconfidence issue on OOD data has been revealed that using mismatched BatchNorm statistics—that are estimated on ID data yet blindly applied to the OOD data in testing—can trigger abnormally high unit activations and model output accordingly (Sun et al., 2021b). Therefore, ReAct (Sun et al., 2021b) proposes truncating the high activations, which establishes strong post-hoc detection performance and further boosts the performance of existing scoring functions. Similarly, NMD (Dong et al., 2022) uses the activation means from BatchNorm layers for ID/OOD discrepancy. While ReAct considers activation space, (Sun & Li, 2022) proposes a weight sparsification-based OOD detection framework termed DICE. DICE ranks weights based on a measure of contribution and selectively uses the most salient weights to derive the output for OOD detection. By pruning away noisy signals, DICE provably reduces the output variance for OOD data, resulting in a sharper output distribution and stronger separability from ID data. In a similar vein, ASH (Djurisic et al., 2023) also targets the activation space but adopts a different strategy. It removes a significant portion (e.g., 90%) of an input's feature representations from a late layer based on a top-K criterion, followed by adjusting the remaining activations (e.g., 10%) either by scaling or assigning constant values, yielding surprisingly effective results.

b. Training-Based Methods With the training phase, confidence can be developed via designing a confidence-estimating branch (DeVries & Taylor, 2018) or class (Wang et al., 2021), ensembling with leaving-out strategy (Vyas et al., 2018), adversarial training (Bitterwolf et al., 2020; Chen et al., 2020b; Hein et al., 2019; Choi & Chung, 2020; Chen et al., 2021c), stronger data augmentation (Hein et al., 2019; Thulasidasan et al., 2019; Yun et al., 2019; DeVries & Taylor, 2017; Hendrycks et al., 2019c, 2022c), pretext training (Tack et al., 2020), better uncertainty modeling (Meinke & Hein, 2019; Bibas et al., 2021), input-level manipulation (Liang et al., 2018; Wang et al., 2022b), and utilizing feature or statistics from the intermediate-layer features (Lin et al., 2021; Dong et al., 2022). Especially, to enhance the sensitivity to covariate shift, some methods focus on the hidden representations in the middle layers of neural networks. Generalized ODIN, or G-ODIN (Hsu et al., 2020) extended ODIN (Liang et al., 2018) by using a specialized training objective termed

DeConf-C and choosing hyperparameters such as perturbation magnitude on ID data. Note that we do not categorize G-ODIN as post-hoc method as it requires model retraining. Recent work (Wei et al., 2022) shows that the overconfidence issue can be mitigated through Logit Normalization (LogitNorm), a simple fix to the common cross-entropy loss by enforcing a constant vector norm on the logits in training. Trained with LogitNorm, neural networks produce highly distinguishable confidence scores between in- and out-of-distribution data.

Some works redesign the label space to achieve good OOD detection performance. While commonly used to encode categorical information for classification, the one-hot encoding ignores the inherent relationship among labels. For example, it is unreasonable to have a uniform distance between `dog` and `cat` vs. `dog` and `car`. To this end, several works attempt to use information in the label space for OOD detection. Some works arrange the large semantic space into a hierarchical taxonomy of known classes (Lee et al., 2018c; Huang et al., 2021; Linderman et al., 2023). Under the redesigned label architecture, top-down classification strategy (Lee et al., 2018c; Linderman et al., 2023) and group softmax training (Huang et al., 2021) are demonstrated effective. Another set of works uses word embeddings to automatically construct the label space. In Shalev et al. (2018), the sparse one-hot labels are replaced with several dense word embeddings from different NLP models, forming multiple regression heads for robust training. When testing, the label, which has the minimal distance to all the embedding vectors from different heads, will be considered as the prediction. If the minimal distance crosses above the threshold, the sample would be classified as "novel". Recent works further take the image features from language-image pre-training models (Radford et al., 2021) to better detect novel classes, where the image encoding space also contains rich information from the language space (Fort et al., 2021; Gan, 2021).

3.1.2 Methods with Outlier Exposure

a. Real Outliers Another branch of OOD detection methods makes use of a set of collected OOD samples, or "outlier", during training to help models learn ID/OOD discrepancy. Starting from the concurrent baselines that encourage a flat/high-entropic prediction on given OOD samples (Hendrycks et al., 2019b; Dhamija et al., 2018) and suppressing OOD feature magnitudes (Dhamija et al., 2018), a follow-up work, MCD (Yu & Aizawa, 2019) uses a network with two branches, between which entropy discrepancy is enlarged for OOD training data. Another straightforward approach with outlier exposure spares an extra abstention (or rejection class) and considers all the given OOD samples in this class (Mohseni et al., 2020; Chen et al., 2021c; Thulasidasan et al., 2021). A later work OECC (Papadopoulos et al., 2021) noticed that an extra regularization for confidence calibration introduces additional improvement for OE. To effectively utilize the given, usually massive, OOD samples, some work use outlier mining (Chen et al., 2021c; Ming et al., 2022b) and adversarial resampling (Li & Vasconcelos, 2020) approaches to obtain a compact yet representative set. In cases where the meaningful "near"-OOD images are not available, MixOE (Zhang et al., 2023b) proposes to interpolate between ID and "far"-OOD images to obtain informative outliers for better regularization. Other works consider a more practical scenario where given OOD samples contain ID samples, therefore using pseudo-labeling (Mohseni et al., 2020) or ID filtering methods (Yang et al., 2021) with optimal transport scheme (Lu et al., 2023) to reduce the interference of ID data. In general, OOD detection with outlier exposure can reach a much better performance.

In typical outlier exposure setups, the auxiliary outlier data used during training is assumed to be representative of the true OOD data encountered at test time. However, research shows that the performance can be largely affected by the correlations between given and real OOD samples (Shafaei et al., 2019). Wang et al. (2023c) and Katz-Samuels et al. (2022) both highlight that the discrepancy between surrogate and test-time OOD distributions can hinder the effectiveness of outlier exposure methods. To address the issue, recent work (Katz-Samuels et al., 2022) proposes a novel framework that enables effectively exploiting unlabeled in-the-wild data for OOD detection. Unlabeled wild data is frequently available since it is produced essentially for free whenever deploying an existing classifier in a real-world system. This setting can be viewed as training OOD detectors in their *natural habitats*, which provide a much better match to the true test time distribution than data collected offline. Wang et al. (2023b) propose an approach to craft an augmented set of OOD distributions around the training outliers to mitigate this distribution shift. Accounting for potential shifts in the OOD data distribution, in addition to shifts in the ID data, is an important consideration for developing robust OOD detectors that can handle the complexities of real-world settings.

b. Outlier Data Generation The outlier exposure approaches impose a strong assumption on the availability of OOD training data, which can be infeasible in practice. When no OOD sample is available, some methods attempt to synthesize OOD samples to enable ID/OOD separability. Existing works leverage GANs to generate OOD training samples and force the model predictions to be uniform (Lee et al., 2018a), generate boundary samples in the low-density region (Vernekar et al., 2019), or similarly, high-confidence OOD samples (Sricharan et al., 2018), or using meta-learning the update sample generation (Jeong & Kim, 2020). However, synthesizing images in the high-dimensional pixel space can be difficult to optimize. Recent work VOS (Du

et al., 2022b) proposed synthesizing virtual outliers from the low-likelihood region in the feature space, which is more tractable given lower dimensionality. While VOS (Du et al., 2022b) is a parametric approach that models the feature space as a class-conditional Gaussian distribution, NPOS (Tao et al., 2023) also generates outlier ID data but in a non-parametric approach. Noticing the generated OOD data could be incorrect or irrelevant, DOE (Wang et al., 2023c) synthesizes hard OOD data that leads to worst judgments to train the OOD detector with a min-max learning scheme, and ATOL (Zheng et al., 2023) uses auxiliary task to relieve the mistaken OOD generation. In object detection, (Du et al., 2022a) proposes synthesizing unknown objects from videos in the wild using spatial-temporal unknown distillation.

3.1.3 Gradient-Based Methods

Existing OOD detection approaches primarily rely on the output (Sect. 3.1) or feature space for deriving OOD scores, while overlooking information from the gradient space. ODIN (Liang et al., 2018) first explored using gradient information for OOD detection. In particular, ODIN proposed using input pre-processing by adding small perturbations obtained from the input gradients. The goal of ODIN perturbations is to increase the softmax score of any given input by reinforcing the model's belief in the predicted label. Ultimately the perturbations have been found to create a greater gap between the softmax scores of ID and OOD inputs, thus making them more separable and improving the performance of OOD detection. While ODIN only uses gradients implicitly through input perturbation, recent work proposed GradNorm (Huang et al., 2021) which explicitly derives a scoring function from the gradient space. GradNorm employs the vector norm of gradients, backpropagated from the KL divergence between the softmax output and a uniform probability distribution. A recent research (Igoe et al., 2022) demonstrates that while gradient-based methods are effective, their success does not necessarily depend on gradients, but rather on the magnitude of learned feature embeddings and predicted output distribution.

3.1.4 Bayesian Models

A Bayesian model is a statistical model that implements Bayes' rule to infer all uncertainty within the model (Jaynes, 1986). The most representative method is the Bayesian neural network (Neal, 2012), which draws samples from the posterior distribution of the model via MCMC (Gamerman & Lopes, 2006), Laplace methods (Mackay, 1992; Foong et al., 2020) and variational inference (Peterson & Hartman, 1989), forming the epistemic uncertainty of the model prediction. However, their obvious shortcomings of inaccurate predictions (Wenzel et al., 2020) and high computational costs (Gelman, 2008) prevent them from wide adoption in practice. Recent works attempt several less principled approximations including MC-dropout (Gal & Ghahramani, 2016) and deep ensembles (Dietterich, 2000; Lakshminarayanan et al., 2017; Maddox et al., 2019) for faster and better estimates of uncertainty. These methods are less competitive for OOD uncertainty estimation. Further exploration takes natural-gradient variational inference and enables practical and affordable modern deep learning training while preserving the benefits of Bayesian principles (Osawa et al., 2019). Dirichlet Prior Network (DPN) is also used for OOD detection with an uncertainty modeling of three different sources of uncertainty: model uncertainty, data uncertainty, and distributional uncertainty, and form a line of works (Malinin & Gales, 2018, 2019; Nandy et al., 2020). Recently, the Bayesian hypothesis test has been used to formulate OOD detection, with upweighting method and Hessian approximation for scalability (Kim et al., 2021).

3.1.5 OOD Detection for Foundation Models

Foundation models (Bommasani et al., 2021), notably large-scale vision-language models (Radford et al., 2021), have demonstrated exceptional performance in a variety of downstream tasks. Their success is largely attributed to extensive pre-training on large-scale datasets. Several works (Hendrycks et al., 2019a; Fort et al., 2021; Hendrycks et al., 2020) reveal that well-pretrained models can significantly enhance OOD detection, particularly in challenging scenarios.

However, adapting (tuning) these models for downstream tasks with specific semantic (label) space in the training data remains a challenge, as simple approaches such as linear probing, prompt tuning (Zhou et al., 2022a, b; Jia et al., 2022), and adaptor-style fine-tuning methods (Gao et al., 2023) do not have good results on OOD detection. Dong et al. (2023) establish a comprehensive few-shot OOD detection benchmark, demonstrating the superiority of parameter-efficient fine-tuning strategies over conventional techniques. They propose a novel method called Domain-Specific and General Knowledge Fusion (DSGF) to strengthen fine-tuned features with original pre-trained features, significantly enhancing few-shot OOD detection capabilities. To advance the problem, a thorough investigation (Ming & Li, 2023) examines how fine-tuned vision-language models are performed. Additionally, recent research (Miyai et al., 2023a) highlights the impact of large-scale pretraining data and provides a systematic study on pretraining strategies on OOD detection performance. On a technical front, LoCoOp (Miyai et al., 2023b) introduces OOD regularization to a subset of CLIP's local features identified as OOD, enhancing prompt learning for better ID and OOD differentiation, and LSA (Lu et al., 2023) uses a bidirectional prompt customization mechanism to enhance the image-text alignment.

The strong zero-shot learning capabilities of models like CLIP (Radford et al., 2021) also open avenues for zero-shot OOD detection. This new setting aims to categorize known class samples and detect samples that do not belong to any of the known classes, where known classes are represented solely through textual descriptions or class names, eliminating the need for explicit training on these classes. Addressing this, ZOC (Esmaeilpour et al., 2022) trains a decoder based on CLIP's visual encoder to create candidate labels for OOD detection. While ZOC is computationally intensive and data-demanding, MCM (Ming et al., 2022a) opts for softmax scaling to align visual features with textual concepts for OOD detection. Jiang et al. (2023c) propose NegLabel, which introduces massive negative labels exhibiting significant semantic differences from ID labels, determining whether an image is OOD by comparing its affinity towards ID and negative labels. Bai et al. (2023) construct ID-like outliers using CLIP and propose an ID-like prompt learning framework to identify challenging OOD samples. Nie et al. (2023) propose to learn a set of negative prompts for each class, which are leveraged along with learned positive prompts to measure similarity and dissimilarity in the feature space simultaneously, enabling more accurate detection of OOD samples. A recent advancement, CLIPN (Wang et al., 2023a), innovatively integrates a "no" logic in OOD detection. Utilizing new prompts and a text encoder, along with novel opposite loss functions, CLIPN effectively tackles the challenge of identifying hard-to-distinguish OOD samples. This development marks a significant stride in enhancing the precision of OOD detection in complex scenarios.

3.2 Density-Based Methods

Density-based methods in OOD detection explicitly model the in-distribution with some probabilistic models, and flag test data in low-density regions as OOD. Although OOD detection can be different from AD in that multiple classes exist in the in-distribution, density estimation methods used for AD in Sect. 4.2 can be directly adapted to OOD detection by unifying the ID data as a whole (Zong et al., 2018; Abati et al., 2019; Pidhorskyi et al., 2018; Deecke et al., 2018; Sabokrou et al., 2018). When the ID contains multiple classes, class-conditional Gaussian distribution can explicitly model the in-distribution so that the OOD samples can be identified based on their likelihoods (Lee et al., 2018b). Flow-based methods (Kobyzev et al., 2020; Zisselman & Tamar, 2020; Kingma & Dhariwal, 2018; Van Oord et al., 2016; Jiang et al., 2021a) can also be used for probabilistic modeling. While directly estimating the likelihood seems like a natural approach, some works (Nalisnick et al., 2018; Choi et al., 2018; Kirichenko et al., 2020) find that probabilistic models sometimes assign a higher likelihood for the OOD sample. Several works attempt to solve the problems using likelihood ratio (Ren et al., 2019). Serrà et al. (2020) finds that the likelihood exhibits a strong bias towards the input complexity and proposes a likelihood ratio-based method to compensate for the influence of input complexity. Recent methods turn to new scores such as likelihood regret (Xiao et al., 2020) or an ensemble of multiple density models (Choi et al., 2018). To directly model the density of semantic space, SEM score is used with a simple combination of density estimation in the low-level and high-level space (Yang et al., 2022b). Overall, generative models can be prohibitively challenging to train and optimize, and the performance can often lag behind the classification-based approaches (Sect. 3.1).

3.3 Distance-Based Methods

The basic idea of distance-based methods is that the testing OOD samples should be relatively far away from the centroids or prototypes of in-distribution classes. (Lee et al., 2018b) uses the minimum Mahalanobis distance to all class centroids for detection. A subsequent work splits the images into foreground and background and then calculates the Mahalanobis distance ratio between the two spaces (Ren et al., 2021). In contrast to the parametric approach, recent work (Sun et al., 2022) shows strong promise of non-parametric nearest-neighbor distance for OOD detection. Unlike Mahalanobis, the non-parametric approach does not impose any distributional assumption about the underlying feature space, hence providing stronger simplicity, flexibility, and generality.

For distance functions, some works use cosine similarity between test sample features and class features to determine OOD samples (Techapanurak et al., 2020; Chen et al., 2020c). The one-dimensional subspace spanned by the first singular vector of the training features is shown to be more suitable for cosine similarity-based detection (Zaeemzadeh et al., 2021). Moreover, other works leverage distances with radial basis function kernel (Van Amersfoort et al., 2020), Euclidean distance (Huang et al., 2020a), and geodesic distance (Gomes et al., 2022) between the input's embedding and the class centroids. Apart from calculating the distance between samples and class centroids, the feature norm in the orthogonal complement space of the principal space is shown effective on OOD detection (Wang et al., 2022a). Recent work CIDER (Ming et al., 2023) explores the usability of the embeddings in the hyperspherical space, where inter-class dispersion and inner-class compactness can be encouraged.

3.4 Reconstruction-Based Methods

The core idea of reconstruction-based methods is that the encoder-decoder framework trained on the ID data usually yields different outcomes for ID and OOD samples. The difference in model performance can be utilized as an indicator

for detecting anomalies. For example, reconstruction models that are only trained by ID data cannot well recover the OOD data (Denouden et al., 2018), and therefore the OOD can be identified. While reconstruction-based models with pixel-level comparison seem not a popular solution in OOD detection for its expensive training cost, reconstructing with hidden features is shown as a promising alternative (Zhou, 2022). Rather than reconstructing the entire image, recent work MoodCat (Yang et al., 2022c) masks a random portion of the input image and identifies OOD samples using the quality of the classification-based reconstruction results. READ (Jiang et al., 2023a) combines inconsistencies from a classifier and an autoencoder by transforming the reconstruction error of raw pixels to the latent space of the classifier. MOOD (Li et al., 2023b) shows that masked image modeling for pretraining is beneficial to OOD detection tasks compared to contrastive training and classic classifier training.

3.5 Theoretical Analysis

Early theoretical research on OOD detection (Zhang et al., 2021) delves into the limitations of Deep Generative Models (DGMs) in OOD contexts. This work uncovers a critical flaw where DGMs frequently assign greater probabilities to OOD data compared to training data, attributing this issue primarily to model misestimation rather than the typical set hypothesis. This hypothesis posits that relevant out-distributions might be located in high-likelihood areas of the data distribution. The study concludes that any generalized OOD task must restrict the set of distributions that are considered out-of-distribution, as without any restrictions, the task is impossible. Later work (Morteza & Li, 2022) advances the field by developing a comprehensive analytical framework aimed at enhancing theoretical understanding and practical performance of OOD detection methods in neural networks. Their innovative approach culminates in a novel OOD detection method that surpasses existing techniques in both theoretical robustness and empirical performance.

Another series of studies has been focused on Open-Set Learning (OSL). The seminal work in this domain (Scheirer et al., 2013) conceptualizes open-space risk for recognizing samples from unknown classes. The following research applies extreme value theory to OSL (Jain et al., 2014; Rudd et al., 2017). While probably approximately correct (PAC) theory is applied for OSR (Liu et al., 2018a), their method required test samples during training. Therefore, an investigation of the generalization error bound is conducted and proves the existence of a low-error OSL algorithm under certain assumptions (Fang et al., 2021). Still, under the PAC theory, a later study establishes necessary and sufficient conditions for the learnability of OOD detection in various scenarios (Fang et al., 2022), including cases with overlapping and non-overlapping ID and OOD data. Their work also offers theoretical support for existing OOD detection algorithms and suggests that OOD detection is possible under certain practical conditions.

Despite these theoretical advancements, the field eagerly anticipates further research addressing aspects such as generalization in OOD detection, the explainability of these models, the integration of deep learning theory specific to OOD detection, and the exploration of foundation model theories pertinent to this area.

3.6 Discussion

The field of OOD detection has enjoyed rapid development since its emergence, with a large space of solutions. In the multi-class setting, the problem can be canonical to OSR (Sect. 4.1)—accurately classify test samples from ID within the class space $\mathcal{Y}$, and reject test samples with semantics outside the support of $\mathcal{Y}$. The difference often lies in the evaluation protocol. OSR splits a dataset into two halves: one set as ID and another set as OOD. In contrast, OOD allows a more general and flexible evaluation by considering test samples from different datasets or domains. Moreover, OOD detection encompasses a broader spectrum of learning tasks (e.g.,multi-label classification (Wang et al., 2021), object detection (Du et al., 2022b, a)) and solution space. Apart from the methodology development, theoretical understanding has also received attention in the community (Morteza & Li, 2022), providing provable guarantees and empirical analysis to understand how OOD detection performance changes with respect to data distributions.

4 Methodologies from Other Sub-tasks

In this section, we briefly introduce methodologies for subtasks under the generalized OOD detection framework, including AD, ND, OSR, and OD, in hope that the methods from other sub-tasks can inspire more ideas for OOD detection community.

4.1 Open Set Recognition

The concept of OSR was first introduced in (Scheirer et al., 2013), which showed the validity of 1-class SVM and binary SVM for solving the OSR problem. In particular, (Scheirer et al., 2013) proposes the 1-vs-Set SVM to manage the open-set risk by solving a two-plane optimization problem instead of the classic half-space of a binary linear classifier. This paper highlighted that the open-set space should also be bounded, in addition to bounding the ID risk.

Classification-Based Methods Early works focused on logits redistribution using the compact abating probability (CAP) (Scheirer et al., 2014) and extreme value theory

(EVT) (Smith, 1990; Castillo, 2012; Jain et al., 2014). In particular, classic probabilistic models lack the consideration of open-set space. CAP explicitly models the probability of class membership abating from ID points to OOD points, and EVT focuses on modeling the tail distribution with extreme high/low values. In the context of deep learning, OpenMax (Bendale & Boult, 2016) first implements EVT for neural networks. OpenMax replaces the softmax layer with an OpenMax layer, which calibrates the logits with a per-class EVT probabilistic model such as Weibull distribution.

To bypass open-set risk construction, some works attained good results without EVT. For example, some work uses a membership loss to encourage high activations for known classes, and uses large-scale external datasets to learn globally negative filters that can reduce the activations of novel images (Perera & Patel, 2019). Apart from explicitly forcing discrepancy between known/unknown classes, other methods extract stronger features through an auxiliary task of transformation classification (Perera et al., 2020), or mutual information maximization between the input image and its latent features (Sun et al., 2021a), etc..

Image generation techniques have been utilized to synthesize unknown samples from known classes, which helps distinguish between known vs. unknown samples (Ge et al., 2017; Neal et al., 2018; Zhou et al., 2021a; Kong & Ramanan, 2021). While these methods are promising on simple images such as handwritten characters, they do not scale to complex natural image datasets due to the difficulty in generating high-quality images in high-dimensional space. Another solution is to successively choose random categories in the training set and treat them as unknown, which helps the classifier to shrink the boundaries and gain the ability to identify unknown classes (Geng & Chen, 2020; Jang & Kim, 2020). Moreover, Schlachter et al. (2019) splits the training data into typical and atypical subsets, which also helps learn compact classification boundaries.

Distance-Based Methods Distance-based methods for OSR require the prototypes to be class-conditional, which allows maintaining the ID classification performance. Category-based clustering and prototyping are performed based on the visual features extracted from the classifiers. OOD samples can be detected by computing the distance *w.r.t.* clusters (Masana et al., 2018; Shu et al., 2020). Some methods also leveraged contrastive learning to learn more compact clusters for known classes (Liu et al., 2020a; Chen et al., 2020a), which enlarge the distance between ID and OOD. CROSR (Yoshihashi et al., 2019) enhances the features by concatenating visual embeddings from both the classifier and reconstruction model for distance computation in the extended feature space. Besides using features from classifiers, GMVAE (Cao et al., 2020) extracts features using a reconstruction VAE, and models the embeddings of the training set as a Gaussian mixture with multiple centroids for the following distance-based operations. Classifiers using nearest neighbors are also adapted for OSR problem (Júnior et al., 2017). By storing the training samples, the nearest neighbor distance ratio is used for identifying unknown samples in testing.

Reconstruction-Based Methods With similar motivations as Sect. 3.4, reconstruction-based methods expect different reconstruction behavior for ID vs. OOD samples. The difference can be captured in the latent feature space or the pixel space of reconstructed images.

By sparsely encoding images from the known classes, open-set samples can be identified based on their dense representation. Techniques such as sparsity concentration index (Zhang & Patel, 2016) and kernel null space methods (Bodesheim et al., 2013; Liu et al., 2017) are used for sparse encoding.

By fixing the visual encoder obtained from standard multi-class training to maintain ID classification performance, C2AE trains a decoder conditioned on label vectors and estimates the reconstructed images using EVT to distinguish unknown classes (Oza & Patel, 2019). Subsequent works use conditional Gaussian distributions by forcing different latent features to approximate class-wise Gaussian models, which enables classifying known samples as well as rejecting unknown samples (Sun et al., 2020). Other methods generate counterfactual images, which help the model focus more on semantics (Yue et al., 2021). Adversarial defense is also considered in Shao et al. (2020) to enhance model robustness.

Discussion Although there is not an independent section for density-based methods, these methods can play an important role and are fused as a critical step in some classification-based methods such as OpenMax (Bendale & Boult, 2016). The density estimation on visual embeddings can effectively detect unknown classes without influencing the classification performance. A hybrid model also uses a flow-based density estimator to detect unknown samples (Zhang et al., 2020).

As introduced in Sect. 2.4, the general goal of OSR and OOD detection is aligned, that is to detect semantic shift from the training data. Therefore, we encourage methods from these two field should learn more from each other. For example, apart from novel methods, OSR research also shows that a good classifier (Vaze et al., 2022b) in the close-set is critical to OSR performance, which should also applicable to OOD detection tasks.

4.2 Anomaly Detection and Novelty Detection

This section reviews methodologies for sensory and semantic AD and one-class ND. Notice that multi-classes ND is covered in the previous. Given homogeneous in-distribution data, approaches include density-based, reconstruction-based,

distance-based, and hybrid methods. We also discuss theoretical works.

Density-Based Methods Density-based methods model normal data (ID) distributions, assuming anomalous test data has low likelihood while normal data has higher likelihood. Techniques include classic density estimation, density estimation with deep generative models, energy-based models, and frequency-based methods.

Parametric density estimation assumes pre-defined distributions (Danuser & Stricker, 1998). Methods involve multivariate Gaussian distribution (De Maesschalck et al., 2000; Leys et al., 2018), mixed Gaussian distribution (Redner & Walker, 1984; Eskin, 2000), and Poisson distribution (Turcotte et al., 2016). Non-parametric density estimation handles more complex scenarios (Izenman, 1991) with histograms (Van Ryzin, 1973; Xie et al., 2012; Kind et al., 2009; Goldstein & Dengel, 2012) and kernel density estimation (KDE) (Parzen, 1962; Desforges et al., 1998; Hu et al., 2018).

Neural networks generate high-quality features to enhance classic density estimation. Techniques include autoencoder (AE) (Kramer, 1991) and variational autoencoder (VAE) (Kingma & Welling, 2013)-based models, generative adversarial networks (GANs) (Goodfellow et al., 2014), flow-based models (Rezende & Mohamed, 2015; Kobyzev et al., 2020), and representation enhancement strategies.

EBMs use scalar energy scores to express probability density (Ngiam et al., 2011) and provide a solution for AD (Zhai et al., 2016). Training EBMs can be computationally expensive, but score matching (Hyvärinen & Dayan, 2005) and stochastic gradient Langevin dynamics (Welling & Teh, 2011) enable efficient training.

Frequency domain analysis for AD includes methods like CNN kernel smoothing (Wang et al., 2020), spectrum-oriented data augmentation (Chen et al., 2021a), and phase spectrum targeting (Liu et al., 2021). These mainly focus on sensory AD.

Reconstruction-Based Methods These AD methods leverage model performance differences on normal and abnormal data in feature space or by reconstruction error.

Sparse reconstruction assumes normal samples can be accurately reconstructed using a limited set of basis functions, while anomalies have larger reconstruction costs and a dense representation (Adler et al., 2015; Li et al., 2017a; Mo et al., 2013). Techniques include L_1 norm-based kernel PCA (Xiao et al., 2013) and low-rank embedded networks (Jiang et al., 2021b).

Reconstruction-error methods assume a model trained on normal data will produce better reconstructions for normal test samples than anomalies. Deep models include AEs (Chen et al., 2018), VAEs (An & Cho, 2015), GANs (Zenati et al., 2018), and U-Net (Liu et al., 2018b).

AE/VAE-based models combine reconstruction-error with AE/VAE models (Chen et al., 2018; An & Cho, 2015) and use strategies like reconstructing by memorized normality (Gong et al., 2019; Park et al., 2020), adapting model architectures (Lai et al., 2020), and partial/conditional reconstruction (Yan et al., 2021; Pidhorskyi et al., 2018; Nguyen et al., 2019). In semi-supervised AD, CoRA (Tian et al., 2019) trains two AEs on inliers and outliers, using reconstruction errors for anomaly detection. Reconstruction-error methods using GANs leverage the discriminator to calculate reconstruction error for anomaly detection (Zenati et al., 2018). Variants like denoising GANs (Sabokrou et al., 2018), class-conditional GANs (Perera et al., 2019), and ensembling (Han et al., 2020) further improve performance. Gradient-based methods observe different patterns on training gradient between normalities and anomalies in a reconstruction task, using gradient-based representation to characterize anomalies (Kwon et al., 2020).

Distance-Based Methods These methods detect anomalies by calculating the distance between samples and prototypes (Wettschereck, 1994), requiring training data in memory. Methods include K-nearest Neighbors (Tian et al., 2014) and prototype-based methods (Münz et al., 2007; Syarif et al., 2012).

Classification-Based Methods AD and one-class ND are often formulated as unsupervised learning problems, but there are some supervised and semi-supervised methods as well. One-class classification (OCC) directly learns a decision boundary that corresponds to a desired density level set of the normal data distribution (Tax, 2002). DeepSVDD (Ruff et al., 2018) introduced classic OCC to the deep learning community. PU learning (Zhang & Zuo, 2008; Bekker & Davis, 2020; Jaskie & Spanias, 2019; Ruff et al., 2020) is proposed for the semi-supervised AD setting where unlabeled data is available in addition to the normal data. Self-supervised learning methods use pretext tasks such as contrastive learning (Tack et al., 2020), image transformation prediction (Bergman & Hoshen, 2020; Golan & El-Yaniv, 2018), and future frame prediction (Georgescu et al., 2021), where anomalies are more likely to make mistakes on the designed task.

One-class classification learns a decision boundary that corresponds to a desired density level set of the normal data distribution, which DeepSVDD (Ruff et al., 2018) introduced to the deep learning community. PU learning (Zhang & Zuo, 2008; Bekker & Davis, 2020; Jaskie & Spanias, 2019; Ruff et al., 2020) is a popular method for the semi-supervised AD setting. Self-supervised learning methods use pretext tasks such as contrastive learning (Tack et al., 2020), image transformation prediction (Bergman & Hoshen, 2020; Golan & El-Yaniv, 2018), and future frame prediction (Georgescu et al., 2021), where anomalies are more likely to make mistakes on the designed task.

Discussion: Sensory *vs* Semantic AD Sensory and semantic AD approaches assume the normal data as homogeneous, despite the presence of multiple categories within it. While semantic AD methods are mainly applicable to sensory AD problems, the latter can benefit from techniques that focus on lower-level features (e.g., flow-based and hidden feature-based), local representations, and frequency-based methods. Although current OOD detection tasks mostly focus on semantic shift, the method for Sensory AD might be especially helpful for far OOD detection, like ImageNet vs Texture dataset.

Discussion: Theoretical Analysis In addition to algorithmic development, theoretical analysis of AD and one-class ND has also been provided in some works. For instance, (Liu et al., 2018a) constructs a clean set of ID and a mixed set of ID/OOD with identical sample sizes, achieving a PAC-style finite sample guarantee for detecting a certain portion of anomalies with the minimum number of false alarms. All these works could be beneficial to the theoretical works of OOD detection.

4.3 Outlier Detection

Outlier detection (OD) observes all samples to identify significant deviations from the majority distribution. Though mostly studied in data mining, deep learning-based OD methods are used for data cleaning in open-set noisy data (Wang et al., 2018; Chen & Gupta, 2015) and open-set semi-supervised learning (Cao et al., 2021).

Density-Based Methods OD methods include Gaussian distribution (Altman & Bland, 2005; Leys et al., 2013), Mahalanobis distance (De Maesschalck et al., 2000), Gaussian mixtures (Yang et al., 2009), and Local outlier factor (LOF) (Breunig et al., 2000). RANSAC (Fischler & Bolles, 1981) estimates parameters for a mathematical model. Classic density methods and NN-based density methods can also be applied.

Distance-Based Methods Outliers can be detected by neighbor counting (Sugiyama & Borgwardt, 2013; Orair et al., 2010), DBSCAN clustering (Ester et al., 1996), and graph-based methods (Hautamaki et al., 2004; Muhlenbach et al., 2004; Liu et al., 2010; Akoglu et al., 2015; Noble & Cook, 2003; Kou et al., 2007; Mingqiang et al., 2012; Wu et al., 2021; Yang et al., 2020a).

Classification-Based Methods AD methods like Isolation Forest (Liu et al., 2008) and OC-SVM (Tax, 2002; Ruff et al., 2018) can be applied to OD. Deep learning models can identify outliers (Li et al., 2017). Techniques for robustness and feature generalizability include ensembling (Nguyen et al., 2020), co-training (Han et al., 2018), and distillation (Li et al., 2017; Yang et al., 2020b).

Discussion OD techniques are valuable for open-set semi-supervised learning, learning with open-set noisy labels, and novelty discovery. All these solutions can be applied especially when OOD samples are exposed during the training stage (Yang et al., 2021).

5 Benchmarks and Experiments

In this section, we report the fair comparison of methodologies that from different categories on the CIFAR (Krizhevsky et al., 2009) benchmark. The report originated from OpenOOD benchmarks (Yang et al., 2022a). We selected several popular AD methods, OOD detection methods (post-hot, training-required, and extra-data-required), and model robustness methods.

5.1 Benchmarks and Metrics

The common practice for building OOD detection benchmarks is to consider an entire dataset as in-distribution (ID), and then collect several datasets that are disconnected from any ID categories as OOD datasets. In this part, we show the results from two popular OOD benchmarks with ID datasets of CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2009) from OpenOOD (c.f.Fig. 4), with each benchmark designing near-OOD and far-OOD datasets to facilitate detailed analysis of the OOD detectors. Near-OOD datasets only have semantic shift compared with ID datasets, while far-OOD further contains obvious covariate (domain) shift.

CIFAR-10 CIFAR-10 (Krizhevsky et al., 2009) is a 10-class dataset for general object classification, which contains 50k training images and 10k test images. As for the OOD dataset, we construct near-OOD with CIFAR-100 (Krizhevsky et al., 2009) and TinyImageNet (Krizhevsky et al., 2012). Notice that 1,207 images are removed from TinyImageNet since they actually belong to CIFAR-10 classes (Yang et al., 2021). Far-OOD is built by MNIST (LeCun & Cortes, 2005), SVHN (Netzer et al., 2011), Texture (Kylberg, 2011), and Places365 (Zhou et al., 2017) with 1,305 images are removed due to semantic overlaps.

CIFAR-100 Another OOD detection benchmark uses CIFAR-100 (Krizhevsky et al., 2009) as an in-distribution, which contains 50k training images and 10k test images with 100 classes. For OOD dataset, near-OOD includes CIFAR-10 (Krizhevsky et al., 2009) and TinyImageNet (Torralba et al., 2008). Similar to the CIFAR-10 benchmark, 2,502 images are removed from TinyImageNet due to the overlapping semantics with CIFAR-100 classes (Yang et al., 2021). Far-OOD consists of MNIST (LeCun & Cortes, 2005), SVHN (Netzer et al., 2011), Texture (Kylberg, 2011), and Places365 (Zhou et al., 2017) with 1,305 images removed.

Metrics We only report the AUROC scores, which measure the area under the Receiver Operating Characteristic (ROC) curve.

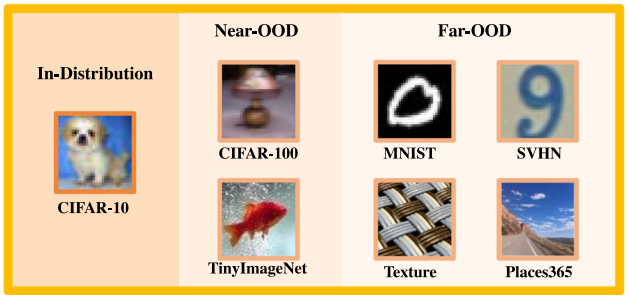

Fig. 4 The illustration of CIFAR-10 benchmark that is used in Sect. 5. The CIFAR-100 benchmark simply swaps the position of CIFAR-10 and CIFAR-100 in the figure

5.2 Experimental Setup

To ensure a fair comparison across methods that originate from different fields and have different implementations, unified settings with common hyperparameters and architecture choices are implemented. ResNet-18 (He et al., 2016) is used as the backbone network. If the implemented method requires training, the widely accepted setting with SGD optimizer, a learning rate of 0.1, momentum of 0.9, and weight decay of 0.0005 for 100 epochs, is used. For further details, please refer to OpenOOD (Yang et al., 2022a; Zhang et al., 2023c).

5.3 Experimental Results and Findings

Data Augmentation Methods are the Most Effective We split Fig. 5 into several sections based on the method type. Generally, the most effective methods are those that use model uncertainty works with data augmentation techniques. This group mainly includes simple and effective methods such as preprocessing methods like PixMix (Hendrycks et al., 2022c) and CutMix (Yun et al., 2019). PixMix achieves 93.1% on Near-OOD in CIFAR-10, the best performance among all the methods in this benchmark. These methods also perform well in most of the other benchmarks. Similarly, other simple and effective methods to enhance model uncertainty estimation such as Ensemble (Dietterich, 2000) and Mixup (Thulasidasan et al., 2019) also demonstrate excellent performance.

Extra Data Seems Not Necessary? Comparing UDG (Yang et al., 2021) (the best from the extra-data part) with KNN (Sun et al., 2022) (the best from the extra data-free part), we found that UDG's advantage is only in CIFAR-10 near-OOD, which is not satisfactory since a large quantity of real outlier data is required. In this benchmark, we use the entire TinyImageNet training set as the extra data, the choice of training outliers could greatly affect the performance of OOD detectors, so further exploration is needed.

Post-Hoc Methods Outperform Training in General Surprisingly, methods that require training do not necessarily perform better. In general, inference-only methods outperform trained methods. Nevertheless, the trained models can be generally used in conjunction with post-hoc methods, which could potentially further increase their performance.

Post-Hoc Methods are Making Progress In general, recent post-hoc methods have had better performance than previous methods since 2021, indicating that the direction of inference-only methods is promising and making progress. Recent methods show improvements in performance on more realistic datasets than previous methods, which focused on toy datasets. For example, the classic MDS performs well on MNIST but poorly on CIFAR-10 and CIFAR-100, while the recent KNN maintains good performance on MNIST, CIFAR-10, CIFAR-100, and also shows outstanding performance on ImageNet (Yang et al., 2022a).

Some AD Methods are Good at Far-OOD Although anomaly detection (AD) methods were originally designed to detect pixel-level appearance differences on the MVTec-AD dataset, they have shown potency in far-OOD detection, such as with DRAEM and CutPaste. Both methods achieved high performance on far-OOD detection, especially when using CIFAR-100 as the in-distribution dataset.

Explore OpenOOD for More Experimental Findings Accompanying our survey, we lead the development of OpenOOD (Yang et al., 2022a), an open-source codebase that provides a unified framework and benchmarking platform for conducting fair comparisons of various model architectures and OOD detection methods. OpenOOD is continuously updated and includes two comprehensive experimental reports (Yang et al., 2022a; Zhang et al., 2023c) that delve into extensive analysis and discovery.[2] We encourage readers to explore OpenOOD's resources for a deeper understanding of key aspects such as selecting model architectures, utilizing pre-trained models, practical applications, and detailed implementation insights.

5.4 Exclusion of Covariate-Shift Detection

While OpenOOD does not include settings for pure covariate shift, this was a deliberate choice. The primary focus is on semantic shifts, which are fundamental to OOD detection. By not separately analyzing covariate shifts, we aim to avoid potential misinterpretations and prevent the overemphasis on covariate shift detection. Experiments in Yang et al. (2022b) highlight a key finding: most current OOD detectors are more sensitive to covariate shifts than semantic shifts and lead to the concept of "full-spectrum OOD detection", advocating for models that effectively generalize to handle covariate shifts while simultaneously detecting samples with

[2] OpenOOD provides a https://zjysteven.github.io/OpenOOD/leaderboard to track SOTAs.

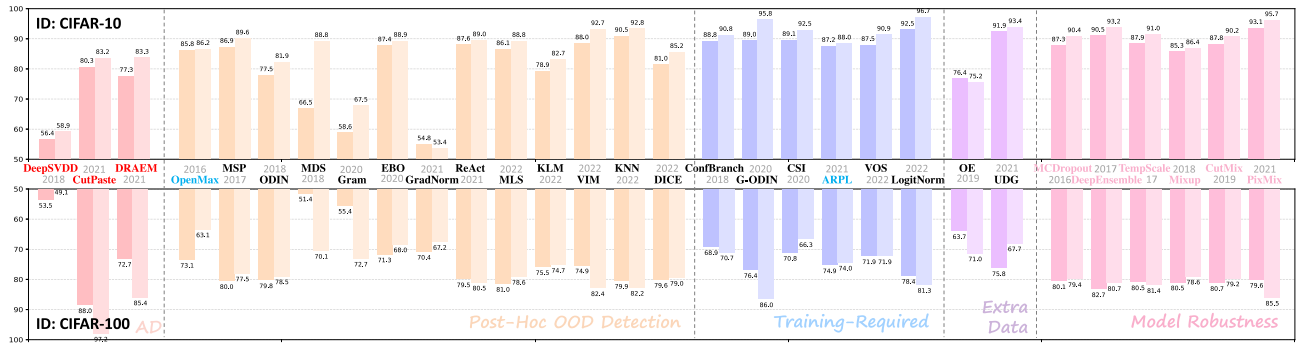

Fig. 5 Comparison between different methodologies under generalized OOD detection framework on the CIFAR-10/100 benchmarks. Results are from OpenOOD (Yang et al., 2022a). Different colors denote the method categories. Each method reports near-OOD (left-bar) and far-OOD (right-bar) AUROC scores, as introduced in Sect. 5.1. Method names in black originated for OOD detection, while in red are AD methods, blue for OSR methods, and pink for models from model uncertainty works (Color figure online)

semantic shifts. More experimental evaluations can be found in OpenOOD v1.5 (Zhang et al., 2023c).

6 Challenges and Future Directions

In this section, we discuss the challenges and future directions of generalized OOD detection.

6.1 Challenges

a. Proper Evaluation and Benchmarking We hope this survey can clarify the distinctions and connections of various sub-tasks, and help future works properly identify the target problem and benchmarks within the framework. The mainstream OOD detection works primarily focus on detecting semantic shifts. Admittedly, the field of OOD detection can be very broad due to the diverse nature of distribution shifts. Such a broad OOD definition also leads to some challenges and concerns (Ahmed & Courville, 2020; Gan, 2021), which advocate a clear specification of OOD type in consideration (e.g.,semantic OOD, adversarial OOD, etc.) so that proposed solutions can be more specialized. Besides, the motivation of detecting a certain distribution shift also requires clarification. While rejecting classifying samples with semantic shift is apparent, detecting sensory OOD should be specified to some meaningful scenarios to contextualize the necessity and practical relevance of the task.

We also urge the community to carefully construct the benchmarks and evaluations. It is noticed that early work (Hendrycks & Gimpel, 2017) ignored the fact that some OOD datasets may contain images with ID categories, causing inaccurate performance evaluation. Fortunately, recent OOD detection works (Yang et al., 2021) have realized this flaw and pay special attention to removing ID classes from OOD samples to ensure proper evaluation.

b. Outlier-Free OOD Detection The outlier exposure approach (Hendrycks et al., 2019b) imposes a strong assumption of the availability of OOD training data, which can be difficult to obtain in practice. Moreover, one needs to perform careful de-duplication to ensure that the outlier training data does not contain ID data. These restrictions may lead to inflexible solutions and prevent the adoption of methods in the real world. Going forward, a major challenge for the field is to devise outlier-free learning objectives that are less dependent on auxiliary outlier dataset.

c. Tradeoff Between Classification and OOD Detection In OSR and OOD detection, it is important to achieve the dual objectives simultaneously: one for the ID task (e.g.,image classification), another for the OOD detection task. For a shared network, an inherent trade-off may exist between the two tasks. Promising solutions should strive for both. These two tasks may or may not contradict each other, depending on the methodologies. For example, (Liu et al., 2019) advocated the integration of image classification and open-set recognition so that the model will possess the capability of discriminative recognition on known classes and sensitivity to novel classes at the same time. (Vaze et al., 2022b) also showed that the ability of detecting novel classes can be highly correlated with its accuracy on the closed-set classes. Yang et al. (2021) demonstrated that optimizing for the cluster compactness of ID classes may facilitate both improved classification and distance-based OOD detection performance. Such solutions may be more desirable than ND, which develops a binary OOD detector separately from the classification model, and requires deploying two models.

d. Real-World Benchmarks and Evaluations Current methods in OOD detection are predominantly evaluated on smaller datasets like CIFAR. However, it has been observed that strategies effective on CIFAR may not perform as well on larger datasets like ImageNet, which has a more extensive semantic space. This discrepancy underscores the

importance of conducting OOD detection evaluations in large-scale, real-world settings. Consequently, we recommend future research to focus on benchmarks based on ImageNet for OOD detection (Huang et al., 2021) and to explore large-scale Open Set Recognition (OSR) benchmarks (Vaze et al., 2022b) to fully test the effectiveness of these methods. Additionally, recent research (Bitterwolf et al., 2023) highlights the presence of erroneous samples in ImageNet OOD benchmarks and introduces the corrected NINCO dataset for more accurate evaluations. Furthermore, expanding the scope of benchmarks to encompass real-world scenarios, such as more realistic datasets (Koh et al., 2021; Cultrera et al., 2023), and object-level OOD detection (Du et al., 2022b, a), can provide valuable insights, especially in safety-critical applications like autonomous driving.

6.2 Future Directions

a. Methodologies Across Sub-tasks Due to the inherent connections among different sub-tasks, their solution space can be shared and inspired by each other. For example, the recent emerging density-based OOD detection research (c.f.Sect. 3.2) can draw insights from the density-based AD methods (c.f.Sect. 4.2) that have been around for a long time.
b. OOD Detection & Generalization An open-world classifier should consider two tasks, i.e.,being robust to covariate shift while being aware of the semantic shift. Existing works pursue these two goals independently. Recent work proposes a semantically coherent OOD detection framework (Yang et al., 2021) that encourages detecting semantic OOD samples while being robust to negligible covariate shift. Given the vague definition of OOD, (Ming et al., 2022c) proposed a formalization of OOD detection by explicitly taking into account the separation between invariant features (semantically related) and environmental features (non-semantic). The work highlighted that spurious environmental features in the training set can significantly impact OOD detection, especially when the semantic OOD data contains the spurious feature. Further, full-spectrum OOD detection (Yang et al., 2022b) highlights the effects of "covariate-shifted in-distribution", and show that most of the previous OOD detectors are unfortunately sensitive to covariate shift rather than semantic shift. This setting explicitly promotes the generalization ability of OOD detectors. Recent works on open long-tailed recognition (Liu et al., 2019), open compound domain adaptation (Liu et al., 2020c), open-set domain adaptation (Panareda Busto & Gall, 2017) and open-set domain generalization (Shu et al., 2021) consider the potential existence of open-class samples. Looking ahead, we envision great research opportunities on how OOD detection and OOD generalization can better enable each other (Liu et al., 2019), in terms of both algorithmic design and comprehensive performance evaluation.

c. OOD Detection & Open-Set Noisy Labels Existing methods of learning from open-set noisy labels focus on suppressing the negative effects of noise (Wang et al., 2018; Li et al., 2021). However, the open-set noisy samples can be useful for outlier exposure (c.f. Sect. 3.1.2) (Wu et al., 2021) and potentially benefit OOD detection. With a similar idea, the setting of open-set semi-supervised learning can be promising for OOD detection. We believe the combination of OOD detection and the previous two fields can provide more insights and possibilities.
e. OOD Detection Enhanced with World Knowledge The existing works utilizing foundation models, particularly multi-modal ones such as CLIP (Radford et al., 2021), have significantly enhanced OOD detection performance, as discussed in Sect. 3.1.5. Starting from this, recent advancements have further focused on leveraging the extensive world knowledge encapsulated in Large Language Models (Dai et al., 2023). This approach aligns with the rapid development in multi-modal world models (Yang et al., 2023; Liu et al., 2023; Li et al., 2023a), presenting burgeoning opportunities for further innovation within the OOD detection community.

rks utilizing foundation mod. OOD Detection For Broader Learning Tasks As mentioned in Sect. 3.6, OOD detection encompasses a broader spectrum of learning tasks, including multi-label classification (Wang et al., 2021), object detection (Du et al., 2022b, a), image segmentation (Hendrycks et al., 2022a), time-series prediction (Kaur et al., 2022b), and LiDAR-based 3D object detection (Nguyen, 2022). For the classification task itself, the researchers also extended the OOD detection technique to improve the reliability of zero-shot pretrained models (Esmaeilpour et al., 2022) (e.g.,CLIP). Furthermore, some studies focus on applying OOD detection methods to produce reliable image captions (Shalev et al., 2022). Recent advancements extend OOD detection to continuously adaptive or online learning environments (Wu et al., 2023). Additionally, OOD detection could show promise to address model reliability issues in broader applications, like mitigating hallucination problems in large language models (Zhou et al., 2020).

Building upon the concept of out-of-distribution (OOD) detection, a recent work introduces unsolvable problem detection (UPD) (Miyai et al., 2024) for question-answering models. UPD asks these models to detect and refrain from predicting answers for unexpected or unsolvable input questions, thereby extending the concept of OOD detection to (visual) question-answering settings. This extension aims to revive OOD detection approaches in broader AI tasks, such as those involving vision-language models (VLMs). Such "unsolvable problem" could be extended to broader contexts, such as dangerous robotic manipulation tasks, and ambiguous or toxic missions.

In sum, we hope the integration of OOD detection methods would promise to enhance the reliability and practicality of

models across various fields, and insights from these fields could, in turn, further refine OOD detection techniques.

7 Conclusion

In this survey, we comprehensively review five topics: AD, ND, OSR, OOD detection, and OD, and unify them as a framework of *generalized OOD detection*. By articulating the motivations and definitions of each sub-task, we encourage follow-up works to accurately locate their target problems and find the most suitable benchmarks. By sorting out the methodologies for each sub-task, we hope that readers can easily grasp the mainstream methods, identify suitable baselines, and contribute future solutions in light of existing ones. By providing insights, challenges, and future directions, we hope that future works will pay more attention to the existing problems and explore more interactions across other tasks within or even outside the scope of generalized OOD detection.

Acknowledgements This study is supported by the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOE-T2EP20221-0012), NTU NAP, and under the RIE2020 Industry Alignment Fund—Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). YL is supported by the Office of the Vice Chancellor for Research and Graduate Education (OVCRGE) with funding from the Wisconsin Alumni Research Foundation (WARF).

Data Availability The datasets analyzed during the current study in Sect. 5 are available in the OpenOOD repository, https://github.com/Jingkang50/OpenOOD.

References

Abati, D., Porrello, A., Calderara, S., & Cucchiara, R. (2019). Latent space autoregression for novelty detection. In *CVPR*.

Adler, A., Elad, M., Hel-Or, Y., & Rivlin, E. (2015). Sparse coding with anomaly detection. *Journal of Signal Processing Systems, 79*, 179–188.

Aggarwal, C. C., & Yu, P. S. (2001). Outlier detection for high dimensional data. In *ACM SIGMOD*.

Ahmed, F., & Courville, A. (2020). Detecting semantic anomalies. In *AAAI*.

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access, 6*, 14410–14430.

Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: A survey. *Data Mining and Knowledge Discovery, 29*, 626–688.

Al-Behadili, H., Grumpe, A., & Wöhler, C. (2015). Incremental learning and novelty detection of gestures in a multi-class system. In *AIMS*.

Altman, D. G., & Bland, J. M. (2005). Standard deviations and standard errors. *BMJ, 6*, 66.

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*, arXiv preprint arXiv:1606.06565

An, J., & Cho, S. (2015). Variational autoencoder based anomaly detection using reconstruction probability. In *Special lecture on IE*.

Angelopoulos, A. N., & Bates, S. (2021). *A gentle introduction to conformal prediction and distribution-free uncertainty quantification*, arXiv preprint arXiv:2107.07511

Atha, D. J., & Jahanshahi, M. R. (2018). Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection. *Structural Health Monitoring, 17*, 1110–1128.

Averly, R., & Chao, W.-L. (2023). *Unified out-of-distribution detection: A model-specific perspective*, arXiv preprint arXiv:2304.06813

Bai, Y., Han, Z., Zhang, C., Cao, B., Jiang, X., & Hu, Q. (2023). *Id-like prompt learning for few-shot out-of-distribution detection*, arXiv preprint arXiv:2311.15243

Bartlett, P. L., & Wegkamp, M. H. (2008). Classification with a reject option using a hinge loss. *Journal of Machine Learning Research, 9*, 8.

Basu, S., & Meckesheimer, M. (2007). Automatic outlier detection for time series: An application to sensor data. *Knowledge and Information Systems, 11*, 137–154.

Bekker, J., & Davis, J. (2020). Learning from positive and unlabeled data: A survey. *Machine Learning, 109*, 719–760.

Bendale, A., & Boult, T. (2015). Towards open world recognition. In *CVPR*.

Bendale, A., & Boult, T. E. (2016). Towards open set deep networks. In *CVPR*.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning, 79*, 151–175.

Ben-Gal, I. (2005). Outlier detection. In *Data mining and knowledge discovery handbook*.

Bergman, L., & Hoshen, Y. (2020). Classification-based anomaly detection for general data. In *ICLR*.

Bergmann, P., Fauser, M., Sattlegger, D., & Steger, C. (2019). Mvtec ad—A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*.

Bianchini, M., Belahcen, A., & Scarselli, F. (2016). A comparative study of inductive and transductive learning with feedforward neural networks. In *Conference of the Italian Association for artificial intelligence*.

Bibas, K., Feder, M., & Hassner, T. (2021). Single layer predictive normalized maximum likelihood for out-of-distribution detection. In *NeurIPS*.

Bitterwolf, J., Meinke, A., & Hein, M. (2020). Certifiably adversarially robust detection of out-of-distribution data. In *NeurIPS*.

Bitterwolf, J., Müller, M., & Hein, M. (2023). In or out? fixing imagenet out-of-distribution detection evaluation. In *ICML*.

Bodesheim, P., Freytag, A., Rodner, E., Kemmler, M., & Denzler, J. (2013). Kernel null space methods for novelty detection. In *CVPR*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., & Brynjolfsson, E. (2021). *On the opportunities and risks of foundation models*, arXiv preprint arXiv:2108.07258

Boult, T. E., Cruz, S., Dhamija, A. R., Gunther, M., Henrydoss, J., & Scheirer, W. J. (2019). Learning and the unknown: Surveying steps toward open world recognition. In *AAAI*.

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). Lof: identifying density-based local outliers. In *SIGMOD*.

Bulusu, S., Kailkhura, B., Li, B., Varshney, P. K., & Song, D. (2020). Anomalous example detection in deep learning: A survey. *IEEE Access, 8*, 132330–132347.

Cai, F., Ozdagli, A. I., Potteiger, N., & Koutsoukos, X. (2021). Inductive conformal out-of-distribution detection based on adversarial autoencoders. In *2021 IEEE international conference on omni-layer intelligent systems (COINS)* (pp. 1–6). IEEE.

Cao, A., Luo, Y., & Klabjan, D. (2020). Open-set recognition with Gaussian mixture variational autoencoders. In *AAAI*.

Cao, K., Brbic, M., & Leskovec, J. (2021). *Open-world semi-supervised learning*, arXiv preprint arXiv:2102.03526

Castillo, E. (2012). *Extreme value theory in engineering*. Elsevier.

Chalapathy, R., & Chawla, S. (2019). *Deep learning for anomaly detection: A survey*, arXiv preprint arXiv:1901.0340

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR), 41*(3), 1–58.

Chen, G., Peng, P., Ma, L., Li, J., Du, L., & Tian, Y. (2021a). Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *ICCV*.

Chen, G., Qiao, L., Shi, Y., Peng, P., Li, J., Huang, T., Pu, S., & Tian, Y. (2020a). Learning open set network with discriminative reciprocal points. In *ECCV*.

Chen, J., Li, Y., Wu, X., Liang, Y., & Jha, S. (2020b). *Robust out-of-distribution detection for neural networks*, arXiv preprint arXiv:2003.09711

Chen, J., Li, Y., Wu, X., Liang, Y., & Jha, S. (2021c). Atom: Robustifying out-of-distribution detection using outlier mining. In *ECML&PKDD*.

Chen, X., & Gupta, A. (2015). Webly supervised learning of convolutional networks. In *ICCV*.

Chen, X., Lan, X., Sun, F., & Zheng, N. (2020c). A boundary based out-of-distribution classifier for generalized zero-shot learning. In *ECCV*.

Chen, Z., Yeo, C. K., Lee, B. S., & Lau, C. T. (2018). Autoencoder-based network anomaly detection. In *Wireless telecommunications symposium*.

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

Choi, S., & Chung, S.-Y. (2020). Novelty detection via blurring. In *ICLR*.

Chow, C. (1970). On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory, 16*, 41–6.

Chu, W.-H., & Kitani, K. M. (2020). Neural batch sampling with reinforcement learning for semi-supervised anomaly detection. In *ECCV*.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*, 273–97.

Cultrera, L., Seidenari, L., & Del Bimbo, A. (2023). Leveraging visual attention for out-of-distribution detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4447–4456).

Dai, Y., Lang, H., Zeng, K., Huang, F., & Li, Y., (2023). *Exploring large language models for multi-modal out-of-distribution detection*, arXiv preprint arXiv:2310.08027

Danuser, G., & Stricker, M. (1998). Parametric model fitting: From inlier characterization to outlier detection. In *TPAMI*.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). *The Mahalanobis distance, chemometrics and intelligent laboratory systems*.

Deecke, L., Vandermeulen, R., Ruff, L., Mandt, S., & Kloft, M. (2018). Image anomaly detection with generative adversarial networks. In *ECML&KDD*.

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

Desforges, M., Jacob, P., & Cooper, J. (1998). Applications of probability density estimation to the detection of abnormal conditions in engineering. In *Proceedings of the institution of mechanical engineers*.

DeVries, T., & Taylor, G. W. (2017). *Improved regularization of convolutional neural networks with cutout*, arXiv preprint arXiv:1708.04552

DeVries, T., & Taylor, G. W. (2018). *Learning confidence for out-of-distribution detection in neural networks*, arXiv preprint arXiv:1802.04865

Dhamija, A. R., Günther, M., & Boult, T. E. (2018). Reducing network agnostophobia. In *NeurIPS*.

Diehl, C. P., & Hampshire, J. B. (2002). Real-time object classification and novelty detection for collaborative video surveillance. In *IJCNN*.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *International workshop on multiple classifier systems*.

Djurisic, A., Bozanic, N., Ashok, A., & Liu, R. (2023). Extremely simple activation shaping for out-of-distribution detection. In *ICLR*.

Dolhansky, B., Howes, R., Pflaum, B., Baram, N., & Ferrer, C. C. (2019). *The deepfake detection challenge (dfdc) preview dataset*, arXiv preprint arXiv:1910.08854

Dong, J., Gao, Y., Zhou, H., Cen, J., Yao, Y., Yoon, S., & Sun, P. D. (2023). *Towards few-shot out-of-distribution detection*, arXiv preprint arXiv:2311.12076

Dong, X., Guo, J., Ang Li, W.-T.T., Liu, C., & Kung, H. (2022a). Neural mean discrepancy for efficient out-of-distribution detection. In *CVPR*.

Dong, X., Guo, J., Li, A., Ting, W.-T., Liu, C., & Kung, H. (2022a). Neural mean discrepancy for efficient out-of-distribution detection. In *CVPR*.

Dou, Y., Li, W., Liu, Z., Dong, Z., Luo, J., & Philip, S. Y. (2019). Uncovering download fraud activities in mobile app markets. In *ASONAM*.

Drummond, N., & Shearer, R. (2006). The open world assumption. In *eSI workshop*.

Du, X., Wang, X., Gozum, G., & Li, Y. (2022a). Unknown-aware object detection: Learning what you don't know from videos in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Du, X., Wang, Z., Cai, M., & Li, Y. (2022b). Vos: Learning what you don't know by virtual outlier synthesis. In *Proceedings of the international conference on learning representations*.

Eskin, E. (2000). Anomaly detection over noisy data using learned probability distributions. In *ICML*.

Esmaeilpour, S., Liu, B., Robertson, E., & Shu, L. (2022). Zero-shot out-of-distribution detection based on the pretrained model clip. In *AAAI*.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*.

Fang, Z., Li, Y., Lu, J., Dong, J., Han, B., & Liu, F. (2022). Is out-of-distribution detection learnable? In *NeurIPS*.

Fang, Z., Lu, J., Liu, A., Liu, F., & Zhang, G. (2021). Learning bounds for open-set learning. In *ICML*.

Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters, 27*, 861–74.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM, 24*, 381–395.

Foong, A. Y., Li, Y., Hernández-Lobato, J. M., & Turner, R. E. (2020). 'in-between' uncertainty in Bayesian neural networks. In *ICML-W*.

Fort, S., Ren, J., & Lakshminarayanan, B. (2021). Exploring the limits of out-of-distribution detection. In *NeurIPS*.

Fumera, G., & Roli, F. (2002). Support vector machines with embedded reject option. In *International workshop on support vector machines*.

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*.

Gamerman, D., & Lopes, H. F. (2006). *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*. CRC Press.

Gan, W. (2021). *Language guided out-of-distribution detection*.

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., & Qiao, Y. (2023). Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision, 132*, 1–15.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *CVPR*.

Ge, Z., Demyanov, S., Chen, Z., & Garnavi, R. (2017). Generative openmax for multi-class open set classification. In *BMVC*.

Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The Kitti vision benchmark suite. In *CVPR*.

Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis, 66*, 4445–449.

Geng, C., & Chen, S. (2020). Collective decision for open set recognition. In *TKDE*.

Geng, C., Huang, S., & Chen, S. (2020). Recent advances in open set recognition: A survey. In *TPAMI*.

Georgescu, M.-I., Barbalau, A., Ionescu, R. T., Khan, F. S., Popescu, M., & Shah, M. (2021). Anomaly detection in video via self-supervised and multi-task learning. In *CVPR*.

Golan, I., & El-Yaniv, R. (2018). Deep anomaly detection using geometric transformations. In *NeurIPS*.

Goldstein, M., & Dengel, A. (2012). Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. In *KI-2012: Poster and demo track*.

Gomes, E. D. C., Alberge, F., Duhamel, P., & Piantanida, P. (2022). Igeood: An information geometry approach to out-of-distribution detection. In *ICLR*.

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., & Hengel, A. V. D. (2019). Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *CVPR*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *NIPS*.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In *ICLR*.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., & Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NIPS*.

Han, K., Vedaldi, A., & Zisserman, A. (2019). Learning to discover novel visual categories via deep transfer clustering. In *CVPR*.

Han, X., Chen, X., & Liu, L.-P. (2020). *Gan ensemble for anomaly detection*, arXiv preprint arXiv:2012.07988

Hautamaki, V., Karkkainen, I., & Franti, P. (2004). Outlier detection using k-nearest neighbour graph. In *ICPR*.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.

Hein, M., Andriushchenko, M., & Bitterwolf, J. (2019). Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*.

Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., & Song, D. (2022a) Scaling out-of-distribution detection for real-world settings. In *ICML*.

Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2021). *Unsolved problems in ML safety*. arXiv preprint, arXiv:2109.13916

Hendrycks, D., & Gimpel, K. (2017). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*.

Hendrycks, D., Lee, K., & Mazeika, M. (2019a). Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning* (pp. 2712–2721). PMLR.

Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., & Song, D. (2020). *Pretrained transformers improve out-of-distribution robustness*, arXiv preprint arXiv:2004.06100

Hendrycks, D., & Mazeika, M. (2022). *X-risk analysis for AI research*. arXiv preprint, arXiv:2206.05862

Hendrycks, D., Mazeika, M., & Dietterich, T. (2019b). Deep anomaly detection with outlier exposure. In *ICLR*.

Hendrycks, D., Mu, N., Cubuk, E. D., Zoph, B., Gilmer, J., & Lakshminarayanan, B. (2019c). Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781

Hendrycks, D., Zou, A., Mazeika, M., Tang, L., Song, D., & Steinhardt, J. (2022c). *Pixmix: Dreamlike pictures comprehensively improve safety measures*.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review, 22*, 85–126.

Hsu, Y.-C., Shen, Y., Jin, H., & Kira, Z. (2020). Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *CVPR*.

Hu, W., Gao, J., Li, B., Wu, O., Du, J., & Maybank, S. (2018). Anomaly detection using local kernel density estimation and context-based regression. In *TKDE*.

Huang, H., Li, Z., Wang, L., Chen, S., Dong, B., & Zhou, X. (2020a). *Feature space singularity for out-of-distribution detection*, arXiv preprint arXiv:2011.14654

Huang, R., Geng, A., & Li, Y. (2021). On the importance of gradients for detecting distributional shifts in the wild. In *NeurIPS*.

Huang, R., & Li, Y. (2021). Mos: Towards scaling out-of-distribution detection for large semantic space. In *CVPR*.

Huang, X., Kroening, D., Ruan, W., Sharp, J., Sun, Y., Thamo, E., Wu, M., & Yi, X. (2020b). A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review, 37*, 100270.

Hyvärinen, A., & Dayan, P. (2005). *Estimation of non-normalized statistical models by score matching*.

Idrees, H., Shah, M., & Surette, R. (2018). Enhancing camera surveillance using computer vision: A research note. *Policing: An International Journal, 41*, 292–307.

Igoe, C., Chung, Y., Char, I., & Schneider, J. (2022). *How useful are gradients for ood detection really?* arXiv preprint arXiv:2205.10439

Izenman, A. J. (1991). Review papers: Recent developments in nonparametric density estimation. *Journal of the American Statistical Association, 86*, 205–224.

Jain, L. P., Scheirer, W. J., & Boult, T. E. (2014). Multi-class open set recognition using probability of inclusion. In *ECCV*.

Jang, J., & Kim, C. O. (2020). *One-vs-rest network-based deep probability model for open set recognition*, arXiv preprint arXiv:2004.08067

Jaskie, K., & Spanias, A. (2019). *Positive and unlabeled learning algorithms and applications: A survey*. In *International conference on information, intelligence, systems and applications*.

Jaynes, E. T. (1986). *Bayesian methods: General background*.

Jeong, T., & Kim, H. (2020). Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification. In *NeurIPS*.

Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., & Lim, S.-N. (2022). Visual prompt tuning. In *European conference on computer vision* (pp. 709–727). Springer.

Jia, X., Han, K., Zhu, Y., & Green, B. (2021). Joint representation learning and novel category discovery on single-and multi-modal data. In *ICCV*.

Jiang, D., Sun, S., & Yu, Y. (2021a). Revisiting flow generative models for out-of-distribution detection. In *International conference on learning representations*.

Jiang, K., Xie, W., Lei, J., Jiang, T., & Li, Y. (2021b). Lren: Low-rank embedded network for sample-free hyperspectral anomaly detection. In *AAAI*.

Jiang, L., Guo, Z., Wu, W., Liu, Z., Liu, Z., Loy, C.C., Yang, S., Xiong, Y., Xia, W., Chen, B., Zhuang, P., Li, S., Chen, S., Yao, T., Ding, S., Li, J., Huang, F., Cao, L., Ji, R., Lu, C., & Tan, G. (2021c). *DeeperForensics Challenge 2020 on real-world face forgery detection: Methods and results*, arXiv preprint arXiv:2102.09471

Jiang, W., Cheng, H., Chen, M., Feng, S., Ge, Y., & Wang, C. (2023a). Read: Aggregating reconstruction error into out-of-distribution detection. In *AAAI*.

Jiang, X., Liu, F., Fang, Z., Chen, H., Liu, T., Zheng, F., & Han, B. (2023b). Detecting out-of-distribution data through in-distribution class prior. In *International conference on machine learning* (pp. 15067–15088). PMLR.

Jiang, X., Liu, F., Fang, Z., Chen, H., Liu, T., Zheng, F., & Han, B. (2023c). Negative label guided ood detection with pretrained vision-language models. In *The twelfth international conference on learning representations*.

Joseph, K., Paul, S., Aggarwal, G., Biswas, S., Rai, P., Han, K., & Balasubramanian, V. N. (2022). Novel class discovery without forgetting. In *ECCV*.

Júnior, P.R.M., De Souza, R. M., Werneck, R. D. O., Stein, B. V., Pazinato, D. V., de Almeida, W. R., Penatti, O. A., Torres, R. D. S., & Rocha, A. (2017). Nearest neighbors distance ratio open-set classifier. *Machine Learning, 6*, 66.

Katz-Samuels, J., Nakhleh, J., Nowak, R., & Li, Y. (2022). Training ood detectors in their natural habitats. In *International conference on machine learning (ICML)*. PMLR.

Kaur, R., Jha, S., Roy, A., Park, S., Dobriban, E., Sokolsky, O., & Lee, I. (2022a). idecode: In-distribution equivariance for conformal out-of-distribution detection. In *Proceedings of the AAAI conference on artificial intelligence* (vol. 36, pp. 7104–7114).

Kaur, R., Sridhar, K., Park, S., Jha, S., Roy, A., Sokolsky, O., & Lee, I. (2022b). *Codit: Conformal out-of-distribution detection in time-series data*, arXiv e-prints.

Kerner, H. R., Wellington, D. F., Wagstaff, K. L., Bell, J. F., Kwan, C., & Amor, H. B. (2019). Novelty detection for multispectral images with application to planetary exploration. In *AAAI*.

Kim, J.-H., Yun, S., & Song, H. O. (2023). Neural relation graph: A unified framework for identifying label noise and outlier data. In *Thirty-seventh conference on neural information processing systems*.

Kim, K., Shin, J., & Kim, H. (2021). Locally most powerful Bayesian test for out-of-distribution detection using deep generative models. In *NeurIPS*.

Kind, A., Stoecklin, M. P., & Dimitropoulos, X. (2009). Histogram-based traffic anomaly detection. *IEEE Transactions on Network and Service Management, 6*, 110–121.

Kingma, D. P., & Dhariwal, P. (2018). Glow: Generative flow with invertible 1x1 convolutions. In *NeurIPS*.

Kingma, D. P., & Welling, M. (2013). *Auto-encoding variational Bayes*, arXiv preprint arXiv:1312.6114

Kirichenko, P., Izmailov, P., & Wilson, A. G. (2020). Why normalizing flows fail to detect out-of-distribution data. in *NeurIPS*.

Kobyzev, I., Prince, S., & Brubaker, M. (2020). Normalizing flows: An introduction and review of current methods. In *TPAMI*.

Koh, P. W., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R. L., Gao, I., & Lee, T. (2021). Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning* (pp. 5637–5664). PMLR.

Kong, S., & Ramanan, D. (2021). Opengan: Open-set recognition via open data generation. In *ICCV*.

Kou, Y., Lu, C.-T., & Dos Santos, R. F. (2007). Spatial outlier detection: A graph-based approach. In *19th IEEE international conference on tools with artificial intelligence (ICTAI)*.

Kramer, M. A. (1991). Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal, 37*, 233–243.

Krizhevsky, A., & Hinton, G. (2009). *Learning multiple layers of features from tiny images*.

Krizhevsky, A., Nair, V., & Hinton, G. (2009). *Cifar-10 and cifar-100 datasets* (vol. 6, (no. 1), p. 1). https://www.cs.toronto.edu/kriz/cifar.html

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NIPS*.

Kwon, G., Prabhushankar, M., Temel, D., & AlRegib, G. (2020). Backpropagated gradient representations for anomaly detection. In *ECCV*.

Kylberg, G. (2011). *Kylberg texture dataset v. 1.0*.

Lai, C.-H., Zou, D., & Lerman, G. (2020). Robust subspace recovery layer for unsupervised anomaly detection. In *ICLR*.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*.

LeCun, Y., & Cortes, C. (2005). *The mnist database of handwritten digits*.

Lee, K., Lee, H., Lee, K., & Shin, J. (2018a). *Training confidence-calibrated classifiers for detecting out-of-distribution samples*.

Lee, K., Lee, K., Lee, H., & Shin, J. (2018b). A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*.

Lee, K., Lee, K., Min, K., Zhang, Y., Shin, J., & Lee, H. (2018c). Hierarchical novelty detection for visual object recognition. In *CVPR*.

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology, 74*, 150–156.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766.

Li, A., Miao, Z., Cen, Y., & Cen, Y. (2017a). Anomaly detection using sparse reconstruction in crowded scenes. *Multimedia Tools and Applications, 76*, 26249–26271.

Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., & Liu, Z. (2023a). *Otter: A multi-modal model with in-context instruction tuning*, *arXiv preprint* arXiv:2305.03726

Li, D., Yang, Y., Song, Y.-Z., & Hospedales, T. M. (2017b). Deeper, broader and artier domain generalization. In *ICCV*.

Li, J., Chen, P., Yu, S., He, Z., Liu, S., & Jia, J. (2023b). Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need. In *CVPR*.

Li, J., Xiong, C., & Hoi, S. C. (2021). Mopro: Webly supervised learning with momentum prototypes. In *ICLR*.

Li, L.-J., & Fei-Fei, L. (2010). Optimol: Automatic online picture collection via incremental model learning. In *IJCV*.

Li, Y., & Vasconcelos, N. (2020). Background data resampling for outlier-aware classification. In *CVPR*.

Li, Y., Yang, J., Song, Y., Cao, L., Luo, J., & Li, L.-J. (2017). Learning from noisy labels with distillation. In *CVPR*.

Liang, S., Li, Y., & Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*.

Lin, Z., Roy, S.D., & Li, Y. (2021). Mood: Multi-level out-of-distribution detection. In *CVPR*.

Linderman, R., Zhang, J., Inkawhich, N., Li, H., & Chen, Y. (2023). Fine-grain inference on out-of-distribution data with hierarchical classification. In S. Chandar, R. Pascanu, H. Sedghi, & D. Precup (Eds.) *Proceedings of the 2nd conference on lifelong learning agents* (vol. 232 of Proceedings of Machine Learning Research, pp. 162–183). PMLR.

Liu, B., Kang, H., Li, H., Hua, G., & Vasconcelos, N. (2020a). Few-shot open-set recognition using meta-learning. In *CVPR*.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *ICDM*.

Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). *Visual instruction tuning*, arXiv preprint arXiv:2304.08485

Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., & Yu, N. (2021). Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *CVPR*.

Liu, H., Shah, S., & Jiang, W. (2004). On-line outlier detection and data cleaning. *Computers & Chemical Engineering, 28*, 1635–1647.

Liu, J., Lian, Z., Wang, Y., & Xiao, J. (2017). Incremental kernel null space discriminant analysis for novelty detection. In *CVPR*.

Liu, S., Garrepalli, R., Dietterich, T., Fern, A., & Hendrycks, D. (2018a). Open category detection with pac guarantees. In *ICML*.

Liu, W., He, J., & Chang, S.-F. (2010). Large graph construction for scalable semi-supervised learning. In *ICML*.

Liu, W., Luo, W., Lian, D., & Gao, S. (2018b). Future frame prediction for anomaly detection—A new baseline. In *CVPR*.

Liu, W., Wang, X., Owens, J. D., & Li, Y. (2020b). Energy-based out-of-distribution detection. In *NeurIPS*.

Liu, X., Lochman, Y., & Zach, C. (2023). Gen: Pushing the limits of softmax-based out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23946–23955).

Liu, Z., Miao, Z., Pan, X., Zhan, X., Lin, D., Yu, S. X., & Gong, B. (2020c). Open compound domain adaptation. In *CVPR*.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., & Yu, S. X. (2019). Large-scale long-tailed recognition in an open world. In *CVPR*.

Loureiro, A., Torgo, L., & Soares, C. (2004). Outlier detection using clustering methods: A data cleaning application. In *Proceedings of KDNet symposium on knowledge-based systems*.

Lu, F., Zhu, K., Zhai, W., Zheng, K., & Cao, Y. (2023). Uncertainty-aware optimal transport for semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3282–3291).

Lu, F., Zhu, K., Zheng, K., Zhai, W., & Cao, Y. (2023). *Likelihood-aware semantic alignment for full-spectrum out-of-distribution detection*, arXiv preprint arXiv:2312.01732

Mackay, D. J. C. (1992). *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology.

Maddox, W. J., Izmailov, P., Garipov, T., Vetrov, D. P., & Wilson, A. G. (2019). A simple baseline for Bayesian uncertainty in deep learning. *Advances in Neural Information Processing Systems, 32*, 13153–13164.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks, *ICLR*.

Mahdavi, A., & Carvalho, M. (2021). *A survey on open set recognition*, arXiv preprint arXiv:2109.00893

Malinin, A., & Gales, M. (2018). Predictive uncertainty estimation via prior networks. In *NeurIPS*.

Malinin, A., & Gales, M. (2019). Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness. In *NeurIPS*.

Markou, M., & Singh, S. (2003a). Novelty detection: A review-part 1: Statistical approaches. *Signal Processing, 83*, 2481–97.

Markou, M., & Singh, S. (2003b). Novelty detection: A review-part 2: Neural network based approaches. *Signal Processing, 83*, 2499–2521.

Masana, M., Ruiz, I., Serrat, J., van de Weijer, J., & Lopez, A. M. (2018). Metric learning for novelty and anomaly detection. In *BMVC*.

Meinke, A., & Hein, M. (2019). *Towards neural networks that provably know when they don't know*, arXiv preprint arXiv:1909.12180

Miljković, D. (2010). *Review of novelty detection methods*. In *MIPRO*.

Ming, Y., Cai, Z., Gu, J., Sun, Y., Li, W., & Li, Y. (2022a). Delving into out-of-distribution detection with vision-language representations. *Advances in Neural Information Processing Systems, 35*, 35087–35102.

Ming, Y., Fan, Y., & Li, Y. (2022b). Poem: Out-of-distribution detection with posterior sampling. In *ICML*.

Ming, Y., & Li, Y. (2023). How does fine-tuning impact out-of-distribution detection for vision-language models? In *IJCV*.

Ming, Y., Sun, Y., Dia, O., & Li, Y. (2023). Cider: Exploiting hyperspherical embeddings for out-of-distribution detection. In *ICLR*.

Ming, Y., Yin, H., & Li, Y. (2022c). On the impact of spurious correlation for out-of-distribution detection. In *AAAI*.

Mingqiang, Z., Hui, H., & Qian, W. (2012). A graph-based clustering algorithm for anomaly intrusion detection. In *International conference on Computer Science & Education (ICCSE)*.

Miyai, A., Yang, J., Zhang, J., Ming, Y., Yu, Q., Irie, G., Li, Y., Li, H., Liu, Z., & Aizawa, K. (2024). *Unsolvable problem detection: Evaluating trustworthiness of vision language models*. arXiv preprint, arXiv:2403.20331

Miyai, A., Yu, Q., Irie, G., & Aizawa, K. (2023a). *Can pre-trained networks detect familiar out-of-distribution data?* arXiv preprint arXiv:2310.00847

Miyai, A., Yu, Q., Irie, G., & Aizawa, K. (2023b). *Locoop: Few-shot out-of-distribution detection via prompt learning*, arXiv preprint arXiv:2306.01293

Mo, X., Monga, V., Bala, R., & Fan, Z. (2013). Adaptive sparse representations for video anomaly detection. *IEEE Transactions on Circuits and Systems for Video Technology, 24*(4), 631–45.

Mohseni, S., Pitale, M., Yadawa, J., & Wang, Z. (2020). Self-supervised learning for generalizable out-of-distribution detection. In *AAAI*.

Mohseni, S., Wang, H., Yu, Z., Xiao, C., Wang, Z., & Yadawa, J. (2021). *Practical machine learning safety: A survey and primer*. arXiv preprint, arXiv:2106.04823

Morteza, P., & Li, Y. (2022). Provable guarantees for understanding out-of-distribution detection. In *AAAI*.

Muhlenbach, F., Lallich, S., & Zighed, D. A. (2004). Identifying and handling mislabelled instances. *Journal of Intelligent Information Systems, 22*, 89–109.

Münz, G., Li, S., & Carle, G. (2007). Traffic anomaly detection using k-means clustering. In *GI/ITG workshop MMBnet*.

Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., & Lakshminarayanan, B. (2018). Do deep generative models know what they don't know? In *NeurIPS*.

Nandy, J., Hsu, W., & Lee, M. L. (2020). Towards maximizing the representation gap between in-domain & out-of-distribution examples. In *NeurIPS*.

Neal, L., Olson, M., Fern, X., Wong, W.-K., & Li, F. (2018). Open set learning with counterfactual images. In *ECCV*.

Neal, R. M. (2012). *Bayesian learning for neural networks*.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., & Ng, A. Y. (2011). *Reading digits in natural images with unsupervised feature learning*.

Ngiam, J., Chen, Z., Koh, P. W., & Ng, A. Y. (2011). Learning deep energy models. In *ICML*.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*.

Nguyen, D. T., Lou, Z., Klar, M., & Brox, T. (2019). Anomaly detection with multiple-hypotheses predictions. In *ICML*.

Nguyen, D. T., Mummadi, C. K., Ngo, T. P. N., Nguyen, T. H. P., Beggel, L., & Brox, T. (2020). Self: Learning to filter noisy labels with self-ensembling. In *ICLR*.

Nguyen, V. D. (2022). *Out-of-distribution detection for lidar-based 3d object detection*, Master's thesis, University of Waterloo.

Nie, J., Zhang, Y., Fang, Z., Liu, T., Han, B., & Tian, X. (2023). Out-of-distribution detection with negative prompts. In *The twelfth international conference on learning representations*.

Nixon, K. A., Aimale, V., & Rowe, R. K. (2008). Spoof detection schemes. In *Handbook of biometrics*.

Noble, C. C., & Cook, D. J. (2003). Graph-based anomaly detection. In *SIGKDD*.

Orair, G. H., Teixeira, C. H., Meira, W., Jr., Wang, Y., & Parthasarathy, S. (2010). Distance-based outlier detection: consolidation and renewed bearing. In *Proceedings of the VLDB endowment*.

Osawa, K., Swaroop, S., Jain, A., Eschenhagen, R., Turner, R. E., Yokota, R., & Khan, M. E. (2019). Practical deep learning with Bayesian principles. In *NeurIPS*.

Oza, P., & Patel, V. M. (2019). C2ae: Class conditioned auto-encoder for open-set recognition. In *CVPR*.

Panareda Busto, P., & Gall, J. (2017). Open set domain adaptation. In *ICCV*.

Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2020). *Deep learning for anomaly detection: A review*, arXiv preprint arXiv:2007.02500

Papadopoulos, A.-A., Rajati, M. R., Shaikh, N., & Wang, J. (2021). Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing, 441*, 138–150.

Park, H., Noh, J., & Ham, B. (2020). Learning memory-guided normality for anomaly detection. In *CVPR*.

Park, J., Chai, J. C. L., Yoon, J., & Teoh, A. B. J. (2023a). Understanding the feature norm for out-of-distribution detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1557–1567).

Park, J., Jung, Y. G., & Teoh, A. B. J. (2023b). Nearest neighbor guidance for out-of-distribution detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1686–1695).

Parmar, J., Chouhan, S., Raychoudhury, V., & Rathore, S. (2023). Open-world machine learning: Applications, challenges, and opportunities. *ACM Computing Surveys, 55*(10), 1–37.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics, 33*, 1065–1076.

Patel, K., Han, H., & Jain, A. K. (2016). Secure face unlock: Spoof detection on smartphones. *IEEE Transactions on Information Forensics and Security, 11*, 2268–2283.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *ICML*.

Perera, P., Morariu, V. I., Jain, R., Manjunatha, V., Wigington, C., Ordonez, V., & Patel, V. M. (2020). Generative-discriminative feature representations for open-set recognition. In *CVPR*.

Perera, P., Nallapati, R., & Xiang, B. (2019). Ocgan: One-class novelty detection using gans with constrained latent representations. In *CVPR*.

Perera, P., & Patel, V. M. (2019). Deep transfer learning for multiple class novelty detection. In *CVPR*.

Peterson, C., & Hartman, E. (1989). Explorations of the mean field theory learning algorithm. *Neural Networks, 2*, 475–494.

Pidhorskyi, S., Almohsen, R., Adjeroh, D. A., & Doretto, G. (2018). Generative probabilistic novelty detection with adversarial autoencoders. In *NeurIPS*.

Pimentel, M. A., Clifton, D. A., Clifton, L., & Tarassenko, L. (2014). A review of novelty detection. *Signal Processing, 99*, 215–249.

Pleiss, G., Souza, A., Kim, J., Li, B., & Weinberger, K. Q. (2019). *Neural network out-of-distribution detection for regression tasks*.

Polatkan, G., Jafarpour, S., Brasoveanu, A., Hughes, S., & Daubechies, I. (2009). Detection of forgery in paintings using supervised learning. In *ICIP*.

Powers, D. M. (2020). Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation. In *JMLT*.

Qui nonero-Candela, J., Sugiyama, M., Lawrence, N. D., & Schwaighofer, A. (2009). *Dataset shift in machine learning*. MIT Press.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., & Krueger, G. (2021). Learning transferable visual models from natural language supervision. In *ICML*.

Redner, R. A., & Walker, H. F. (1984). Mixture densities, maximum likelihood and the em algorithm. *SIAM Review, 26*(2), 195–239.

Ren, J., Fort, S., Liu, J., Roy, A. G., Padhy, S., & Lakshminarayanan, B. (2021). *A simple fix to Mahalanobis distance for improving near-ood detection*, arXiv preprint arXiv:2106.09022

Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., DePristo, M. A., Dillon, J. V., & Lakshminarayanan, B. (2019). Likelihood ratios for out-of-distribution detection. In *NeurIPS*.

Rezende, D., & Mohamed, S. (2015). Variational inference with normalizing flows. In *ICML*.

Rudd, E. M., Jain, L. P., Scheirer, W. J., & Boult, T. E. (2017). The extreme value machine. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(3), 762–768.

Ruff, L., Kauffmann, J. R., Vandermeulen, R. A., Montavon, G., Samek, W., Kloft, M., Dietterich, T. G., & Müller, K.-R. (2021). A unifying review of deep and shallow anomaly detection. In *Proceedings of the IEEE*.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., & Kloft, M. (2018). Deep one-class classification. In *ICML*.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, K.-R. Müller, E., & Kloft, M. (2020). Deep semi-supervised anomaly detection. In *ICLR*.

Sabokrou, M., Khalooei, M., Fathy, M., & Adeli, E. (2018). Adversarially learned one-class classifier for novelty detection. In *CVPR*.

Salehi, M., Mirzaei, H., Hendrycks, D., Li, Y., Rohban, M. H., & Sabokrou, M. (2021). *A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges*, arXiv preprint arXiv:2110.14051

Sastry, C. S., & Oore, S. (2019). Detecting out-of-distribution examples with in-distribution examples and gram matrices. In *NeurIPS-W*.

Sastry, C. S., & Oore, S. (2020). Detecting out-of-distribution examples with gram matrices. In *ICML*.

Scheirer, W. J., de Rezende Rocha, A., Sapkota, A., & Boult, T. E. (2013). Toward open set recognition. In *TPAMI*.

Scheirer, W. J., Jain, L. P., & Boult, T. E. (2014). Probability models for open set recognition. In *TPAMI*.

Schlachter, P., Liao, Y., & Yang, B. (2019). Open-set recognition using intra-class splitting. In *EUSIPCO*.

Sedlmeier, A., Gabor, T., Phan, T., Belzner, L., & Linnhoff-Popien, C. (2019). *Uncertainty-based out-of-distribution detection in deep reinforcement learning*, arXiv preprint arXiv:1901.02219

Serrà, J., Álvarez, D., Gómez, V., Slizovskaia, O., Nú nez, J. F., & Luque, J. (2020). *Input complexity and out-of-distribution detection with likelihood-based generative models*.

Shafaei, A., Schmidt, M., & Little, J. J. (2019). A less biased evaluation of out-of-distribution sample detectors. In *BMVC*.

Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research, 9*(3), 66.

Shalev, G., Adi, Y., & Keshet, J. (2018). Out-of-distribution detection using multiple semantic label representations. In *NeurIPS*.

Shalev, G., Shalev, G.-L., & Keshet, J. (2022). *A baseline for detecting out-of-distribution examples in image captioning*. arXiv preprint, arXiv:2207.05418

Shao, R., Perera, P., Yuen, P. C., & Patel, V. M. (2020). Open-set adversarial defense. In *ECCV*.

Shu, Y., Cao, Z., Wang, C., Wang, J., & Long, M. (2021). Open domain generalization with domain-augmented meta-learning. In *CVPR*.

Shu, Y., Shi, Y., Wang, Y., Huang, T., & Tian, Y. (2020). p-odn: Prototype-based open deep network for open set recognition. *Scientific Reports, 10*, 7146.

Smith, R. L. (1990). Extreme value theory. *Handbook of Applicable Mathematics, 7*, 18.

Sorio, E., Bartoli, A., Davanzo, G., & Medvet, E. (2010). Open world classification of printed invoices. In *Proceedings of the 10th ACM symposium on document engineering*.

Sricharan, K., & Srivastava, A. (2018). Building robust classifiers through generation of confident out of distribution examples. In *NeurIPS-W*.

Sugiyama, M., & Borgwardt, K. (2013). Rapid distance-based outlier detection via sampling. In *NIPS*.

Sun, X., Ding, H., Zhang, C., Lin, G., & Ling, K.-V. (2021a). *M2iosr: Maximal mutual information open set recognition*, arXiv preprint arXiv:2108.02373

Sun, X., Yang, Z., Zhang, C., Ling, K.-V., & Peng, G. (2020). Conditional Gaussian distribution learning for open set recognition. In *CVPR*.

Sun, Y., Guo, C., & Li, Y. (2021b). React: Out-of-distribution detection with rectified activations. In *NeurIPS*.

Sun, Y., & Li, Y. (2022). Dice: Leveraging sparsification for out-of-distribution detection. In *ECCV*.

Sun, Y., Ming, Y., Zhu, X., & Li, Y. (2022). Out-of-distribution detection with deep nearest neighbors. In *ICML*.

Syarif, I., Prugel-Bennett, A., & Wills, G. (2012). Unsupervised clustering approach for network anomaly detection. In *International conference on networked digital technologies*.

Tack, J., Mo, S., Jeong, J., & Shin, J. (2020). Csi: Novelty detection via contrastive learning on distributionally shifted instances. In *NeurIPS*.

Tao, L., Du, X., Zhu, X., & Li, Y. (2023). Non-parametric outlier synthesis. In *ICLR*.

Tariq, M. I., Memon, N. A., Ahmed, S., Tayyaba, S., Mushtaq, M. T., Mian, N. A., Imran, M., & Ashraf, M. W. (2020). A review of deep learning security and privacy defensive techniques. *Mobile Information Systems, 2020*, 1–8.

Tax, D. M. J. (2002). *One-class classification: Concept learning in the absence of counter-examples*.

Techapanurak, E., Suganuma, M., & Okatani, T. (2020). Hyperparameter-free out-of-distribution detection using cosine similarity. In *ACCV*.

Thulasidasan, S., Chennupati, G., Bilmes, J., Bhattacharya, T., & Michalak, S. (2019). On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *NeurIPS*.

Thulasidasan, S., Thapa, S., Dhaubhadel, S., Chennupati, G., Bhattacharya, T., & Bilmes, J. (2021). *An effective baseline for robustness to distributional shift*, arXiv preprint arXiv:2105.07107

Tian, J., Azarian, M. H., & Pecht, M. (2014). Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm. In *PHM society European conference*.

Tian, K., Zhou, S., Fan, J., & Guan, J. (2019). Learning competitive and discriminative reconstructions for anomaly detection. In *AAAI*.

Torralba, A., Fergus, R., & Freeman, W. T. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. In *TPAMI*.

Turcotte, M., Moore, J., Heard, N., & McPhall, A. (2016). Poisson factorization for peer-based anomaly detection. In *IEEE conference on intelligence and security informatics (ISI)*.

Van Amersfoort, J., Smith, L., Teh, Y. W., & Gal, Y. (2020). Uncertainty estimation using a single deep deterministic neural network. In *ICML*.

Van den Broeck, J., Argeseanu Cunningham, S., Eeckels, R., & Herbst, K. (2005). Data cleaning: Detecting, diagnosing, and editing data abnormalities. *PLoS Medicine, 2*, 267.

Van Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. In *ICML*.

Van Ryzin, J. (1973). A histogram method of density estimation. *Communications in Statistics-Theory and Methods, 2*, 493–506.

Vaze, S., Han, K., Vedaldi, A., & Zisserman, A. (2022a). Generalized category discovery. In *CVPR*.

Vaze, S., Han, K., Vedaldi, A., & Zisserman, A. (2022b). Open-set recognition: A good closed-set classifier is all you need. In *ICLR*.

Vernekar, S., Gaurav, A., Abdelzad, V., Denouden, T., Salay, R., & Czarnecki, K. (2019). Out-of-distribution detection in classifiers via generation. In *NeurIPS-W*.

Vinyals, O., Ewalds, T., Bartunov, S., Georgiev, P., Vezhnevets, A. S., Yeo, M., Makhzani, A., Küttler, H., Agapiou, J., Schrittwieser, J., & Quan, J. (2017). *Starcraft II: A new challenge for reinforcement learning*, arXiv preprint arXiv:1708.04782

Vyas, A., Jammalamadaka, N., Zhu, X., Das, D., Kaul, B., & Willke, T. L. (2018). Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In *ECCV*.

Wang, H., Bah, M. J., & Hammad, M. (2019a). Progress in outlier detection techniques: A survey. *IEEE Access, 7*, 107964–108000.

Wang, H., Li, Y., Yao, H., & Li, X. (2023a). Clipn for zero-shot ood detection: Teaching clip to say no. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1802–1812).

Wang, H., Li, Z., Feng, L., & Zhang, W. (2022a). Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Wang, H., Liu, W., Bocchieri, A., & Li, Y. (2021). Can multi-label classification networks know what they don't know? *NeurIPS, 34*, 29074–29087.

Wang, H., Wu, X., Huang, Z., & Xing, E. P. (2020). High-frequency component helps explain the generalization of convolutional neural networks. In *CVPR*.

Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing, 312*, 135–153.

Wang, Q., Fang, Z., Zhang, Y., Liu, F., Li, Y., & Han, B. (2023b). Learning to augment distributions for out-of-distribution detection. *Advances in Neural Information Processing Systems, 36*, 66.

Wang, Q., Liu, F., Zhang, Y., Zhang, J., Gong, C., Liu, T., & Han, B. (2022b). Watermarking for out-of-distribution detection. In *NeurIPS*.

Wang, Q., Ye, J., Liu, F., Dai, Q., Kalander, M., Liu, T., Hao, J., & Han, B. (2023c). *Out-of-distribution detection with implicit outlier transformation*.

Wang, W., Zheng, V. W., Yu, H., & Miao, C. (2019b). A survey of zero-shot learning: Settings, methods, and applications. In *TIST*.

Wang, Y., Li, B., Che, T., Zhou, K., Liu, Z., & Li, D. (2021). Energy-based open-world uncertainty modeling for confidence calibration. In *ICCV*.

Wang, Y., Liu, W., Ma, X., Bailey, J., Zha, H., Song, L., & Xia, S.-T. (2018). Iterative learning with open-set noisy labels. In *CVPR*.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., & Li, Y. (2022). Mitigating neural network overconfidence with logit normalization. In *ICML*.

Welling, M., & Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*.

Wen, D., Han, H., & Jain, A. K. (2015). Face spoof detection with image distortion analysis. *IEEE Transactions on Information Forensics and Security, 10*, 746–761.

Wenzel, F., Roth, K., Veeling, B. S., Świątkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., & Nowozin, S. (2020). How good is the Bayes posterior in deep neural networks really? In *ICML*.

Wettschereck, D. (1994). *A study of distance-based machine learning algorithms*.

Wikipedia contributors. (2021). *Outlier from Wikipedia, the free encyclopedia*. Retrieved August 12, 2021

Wu, X., Lu, J., Fang, Z., & Zhang, G. (2023). Meta ood learning for continuously adaptive ood detection. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 19353–19364).

Wu, Z.-F., Wei, T., Jiang, J., Mao, C., Tang, M., & Li, Y.-F. (2021). Ngc: A unified framework for learning with open-world noisy data. In *ICCV*.

Xia, Y., Cao, X., Wen, F., Hua, G., & Sun, J. (2015). Learning discriminative reconstructions for unsupervised outlier removal. In *CVPR*.

Xiao, T., Zhang, C., & Zha, H. (2015). Learning to detect anomalies in surveillance video. *IEEE Signal Processing Letters, 22*, 1477–1481.

Xiao, Y., Wang, H., Xu, W., & Zhou, J. (2013). L1 norm based kpca for novelty detection. *Pattern Recognition, 46*, 389–396.

Xiao, Z., Yan, Q., & Amit, Y. (2020). Likelihood regret: An out-of-distribution detection score for variational auto-encoder. In *NeurIPS*.

Xie, M., Hu, J., & Tian, B. (2012). Histogram-based online anomaly detection in hierarchical wireless sensor networks. In *ICTSPCC*.

Xu, H., Liu, B., Shu, L., & Yu, P. (2019). Open-world learning and application to product classification. In *WWW*.

Yan, X., Zhang, H., Xu, X., Hu, X., & Heng, P.-A. (2021). Learning semantic context from normal samples for unsupervised anomaly detection. In *AAAI*.

Yang, J., Chen, W., Feng, L., Yan, X., Zheng, H., & Zhang, W. (2020a). Webly supervised image classification with metadata: Automatic noisy label correction via visual-semantic graph. In *ACM multimedia*.

Yang, J., Feng, L., Chen, W., Yan, X., Zheng, H., Luo, P., & Zhang, W. (2020b). Webly supervised image classification with self-contained confidence. In *ECCV*.

Yang, J., Wang, H., Feng, L., Yan, X., Zheng, H., Zhang, W., & Liu, Z. (2021). Semantically coherent out-of-distribution detection. In *ICCV*.

Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., Du, X., Zhou, K., Zhang, W., Hendrycks, D., Li, Y., & Liu, Z. (2022a). Openood: Benchmarking generalized out-of-distribution detection. In *NeurIPS*.

Yang, J., Zhou, K., & Liu, Z. (2022b). *Full-spectrum out-of-distribution detection*, arXiv preprint arXiv:2204.05306

Yang, P., Baracchi, D., Ni, R., Zhao, Y., Argenti, F., & Piva, A. (2020c). A survey of deep learning-based source image forensics. *Journal of Imaging, 6*, 66.

Yang, X., Latecki, L. J., & Pokrajac, D. (2009). Outlier detection with globally optimal exemplar-based gmm. In *SIAM*.

Yang, Y., Gao, R., & Xu, Q. (2022c). *Out-of-distribution detection with semantic mismatch under masking*. In *ECCV*.

Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., & Wang, L. (2023). *The dawn of lmms Preliminary explorations with gpt-4v (ision)*. arXiv preprint, arXiv:2309.17421

Yoshihashi, R., Shao, W., Kawakami, R., You, S., Iida, M., & Naemura, T. (2019). Classification-reconstruction learning for open-set recognition. In *CVPR*.

Yu, Q., & Aizawa, K. (2019). Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *ICCV*.

Yue, Z., Wang, T., Sun, Q., Hua, X.-S., & Zhang, H. (2021). Counterfactual zero-shot and open-set visual recognition. In *CVPR*.

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *CVPR*.

Zaeemzadeh, A., Bisagno, N., Sambugaro, Z., Conci, N., Rahnavard, N., & Shah, M. (2021). Out-of-distribution detection using union of 1-dimensional subspaces. In *CVPR*.

Zenati, H., Foo, C. S., Lecouat, B., Manek, G., & Chandrasekhar, V. R. (2018). Efficient gan-based anomaly detection. In *ICLR-W*.

Zhai, S., Cheng, Y., Lu, W., & Zhang, Z. (2016). Deep structured energy based models for anomaly detection. In *ICML*.

Zhang, B., & Zuo, W. (2008). Learning from positive and unlabeled examples: A survey. In *International symposiums on information processing*.

Zhang, H., Li, A., Guo, J., & Guo, Y. (2020). Hybrid models for open set recognition. In *ECCV*.

Zhang, H., & Patel, V. M. (2016). Sparse representation-based open set recognition. In *TPAMI*.

Zhang, J., Fu, Q., Chen, X., Du, L., Li, Z., Wang, G., Han, S., & Zhang, D. (2023a). Out-of-distribution detection based on in-distribution data patterns memorization with modern Hopfield energy. In *ICLR*.

Zhang, J., Inkawhich, N., Linderman, R., Chen, Y., & Li, H. (2023b). Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)* (pp. 5531–5540).

Zhang, J., Yang, J., Wang, P., Wang, H., Lin, Y., Zhang, H., Sun, Y., Du, X., Zhou, K., Zhang, W., Li, Y., Liu, Z., Chen, Y., & Li, H. (2023c). *Openood v1.5: Enhanced benchmark for out-of-distribution detection*. arXiv preprint, arXiv:2306.09301

Zhang, L., Goldstein, M., & Ranganath, R. (2021). Understanding failures in out-of-distribution detection with deep generative models. In *ICML*.

Zhao, B., & Han, K. (2021). Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *NeurIPS*.

Zheng, H., Wang, Q., Fang, Z., Xia, X., Liu, F., Liu, T., & Han, B. (2023). Out-of-distribution detection learning with unreliable out-of-distribution sources. In *NeurIPS*.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*, 1452–1464.

Zhou, C., Neubig, G., Gu, J., Diab, M., Guzman, P., Zettlemoyer, L., & Ghazvininejad, M. (2020). Detecting hallucinated content in conditional neural sequence generation. In *ACL*.

Zhou, D.-W., Ye, H.-J., & Zhan, D.-C. (2021a). Learning placeholders for open-set recognition. In *CVPR*.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2021b). *Domain generalization: A survey*, arXiv preprint arXiv:2103.02503

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022a). Learning to prompt for vision-language models. In *International Journal of Computer Vision (IJCV)*.

Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022b). Conditional prompt learning for vision-language models. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*.

Zhou, Y. (2022). Rethinking reconstruction autoencoder-based out-of-distribution detection. In *CVPR*.

Zimmerer, D., Full, P. M., Isensee, F., Jäger, P., Adler, T., Petersen, J., Köhler, G., Ross, T., Reinke, A., Kascenas, A., & Jensen, B. S. (2022). Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging, 41*, 2728–2738.

Zisselman, E., & Tamar, A. (2020). Deep residual flow for out of distribution detection. In *CVPR*.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., & Chen, H. (2018). Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In *ICLR*.