Generalized Out-of-Distribution Detection: A Survey

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu

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 before and cannot make a safe decision. This problem 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 are closely related to OOD detection in terms of motivation and methodology. These include anomaly detection (AD), novelty detection (ND), open set recognition (OSR), and outlier detection (OD). Despite having different definitions and problem settings, these problems often confuse readers and practitioners, and as a result, some existing studies misuse terms. In this survey, we first present a generic 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. Then, we conduct a thorough review of each of the five areas by summarizing their recent technical developments. We conclude this survey with open challenges and potential research directions.

Index Terms—Anomaly Detection, Novelty Detection, Open Set Recognition, Out-of-Distribution Detection, Outlier Detection

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) [1], [2], [3], [4], [5]. 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.

Most existing machine learning models are trained based on the closed-world assumption [6], [7], 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 [8], 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) [9], or covariate shift (*e.g.*, OOD samples from a different domain) [10], [11], [12].

The detection of semantic distribution shift (*e.g.*, due to the occurrence of new classes) is the focal point of OOD *detection* tasks considered in this paper, 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) [13], [14], [15], [16], anomaly detection (AD) [17], [18], [19], [20], novelty detection (ND) [21], [22], [23], [24], and open set recognition (OSR) [25], [26], [27]. While all these problems

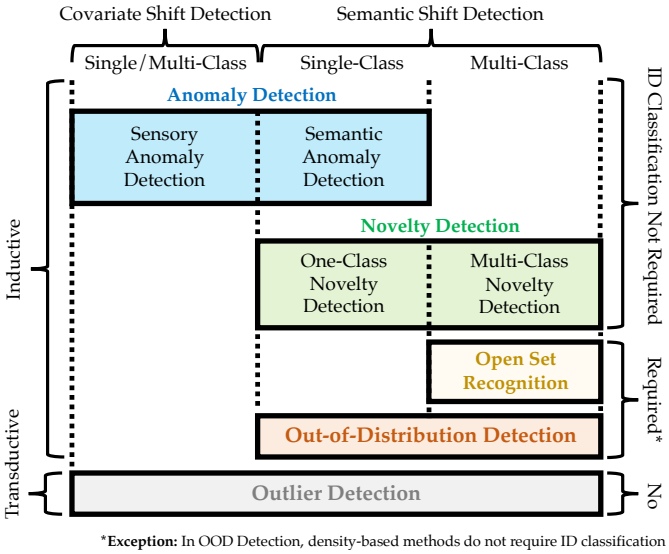


Fig. 1: Taxonomy of generalized OOD detection framework. Four bases are used for the task taxonomy: **(1)** Distribution shift to detect: the task focuses on detecting covariate shift or semantic shift; **(2)** ID data type: the ID data contains one single class or multiple classes; **(3)** Whether the task requires ID classification; **(4)** Transductive learning task requires all observations; inductive tasks follow the train-test scheme. Note that OOD detection can encompass a broader spectrum of learning tasks, and go beyond multi-class classification.

- *J. Yang, K. Zhou, and Z. Liu are with S-Lab, School of Computer Science and Engineering, Nanyang Technological University, Singapore, 639798. E-mail: {jingkang001,kaiyang.zhou,ziwei.liu}@ntu.edu.sg*
- *Y. Li is with Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, United States, 53706. E-mail: sharonli@cs.wisc.edu*

Manuscript submitted October 22, 2021. Issues, comments, and questions are all welcomed in https://github.com/Jingkang50/OODSurvey.

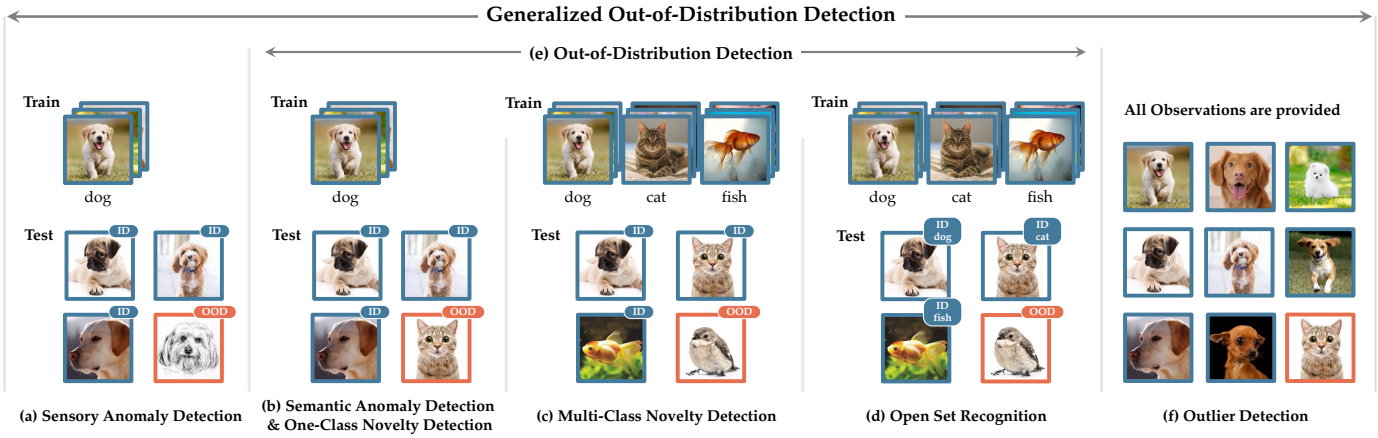


Fig. 2: Exemplar problem settings for tasks under generalized OOD detection framework. Tags on test images refer to model's expected predictions. **(a)** In sensory anomaly detection, test images with covariate shift will be considered as OOD. No semantic shift occurs in this setting. **(b)** In semantic anomaly detection and one-class novelty detection, normality/ID images belong to one class. Test images with semantic shift will be considered as OOD. No covariate shift occurs in this setting. **(c)** In multi-class novelty detection, ID images belong to multiple classes. Test images with semantic shift will be considered as OOD. No covariate shift occurs in this setting. **(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 in-distribution classification. **(e)** Out-of-distribution detection is a super-category that covers semantic AD, one-class ND, multi-class ND, and open-set recognition, which canonically aims to detect test samples with semantic shift without losing the ID classification accuracy. **(f)** 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 samples.

are related to each other by sharing similar motivations, subtle differences exist among the *sub-topics* in terms of the specific definition. However, existing studies often misuse terms and even datasets, due to a lack of comprehensive understanding of the relations among different problems.

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. We further conduct a thorough review of each sub-topic and summarize recent technical developments. To sum up, we make three contributions to the OOD detection 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 more unified framework of *generalized OOD detection*. Under this framework, the similarities and differences of the five sub-topics can be 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**: We conduct a thorough review of the existing methods developed for each sub-topic, with a particular focus on computer vision and deep learning-based approaches. Despite targeting different problem settings, the methods developed within each area can be generally categorized into four groups: 1) density-based methods, 2) reconstruction-based methods, 3) classification-based methods, and 4) distance-based methods. We hope our survey can help readers build a better understanding of the developments for each problem.
3) **Future Research Directions**: Finally, 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.

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 detection (OOD), 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 Section 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. Thereafter, we conclude this section by discussing and articulating the relationships among the sub-topics (Section 2.6). We also note that AD contains 2 sub-tasks of sensory AD and semantic AD, ND contains 2 sub-tasks of one-class ND and multi-class ND. Therefore, we consider a

total of 7 sub-tasks under the framework. Fig. 2 illustrates the settings for each sub-topic.

Preliminary Key to our framework, the notion of distribution shift is very broad and can exhibit in various forms. There are two general types of distribution shift: covariate shift and semantic (label) shift. Formally, let $\mathcal{X}$ and $\mathcal{Y}$ be the input (sensory) and the label (semantic) space, respectively. A data distribution is defined as a joint distribution $P(X, Y)$ over $\mathcal{X} \times \mathcal{Y}$. Distribution shift can occur in either the marginal distribution $P(X)$, $P(Y)$, or both.

Examples of covariate distribution shift on $P(X)$ include adversarial examples [28], [29], domain shift [30], and style changes [31]. Importantly, we note that covariate shifts are 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. 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. For example, to create a "hotdog/not-hotdog detector", we define the concept of `hotdog` to be normal, and those that violate this definition are identified as anomalies, *i.e.*, `not-hotdog`. Note that in this case, `hotdog` is a homogeneous concept, regardless of the possible differences in size, sauce, bread type, price, the origin of the sausage, *etc*. Any other object, such as steaks, rice, and non-food objects like cats and dogs, are all considered as anomalies.

Note that existing anomaly detection settings often restrict the environment of interest to some specific scenarios. For example, the "hotdog/not-hotdog detector" only focuses on realistic images, assuming the nonexistence of images from domains such as cartoons and 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, in anomaly detection settings, the key is to treat normality homogeneously, and detect all possible anomalous samples under some specific scenarios.

Definition Anomaly detection (AD) aims to detect any anomalous samples that are deviated from the predefined normality during testing. The deviation can happen due to either covariate shift or semantic shift, while assuming the other distribution shift do not exist. This leads to two sub-tasks: sensory AD and semantic AD [17], respectively.

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. No covariate shift happens in semantic AD settings.

Two broad categories of anomaly detection techniques exist. In the standard unsupervised AD setting, all given training samples are normal samples. The (semi-)supervised AD setting requires a dataset that has been labeled as "normal" and "abnormal", and involves training a model explicitly.

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 assumes ID only contains samples from a single class or with a homogeneous characteristic.

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.

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 [32], forgery recognition of biometrics and artworks [33], [34], [35], [36], image forensics [37], [38], [39], industrial inspection [40], [41], [42], *etc*. One popular real-world AD benchmark is MVTec [40] for industrial inspection.

In contrast to sensory AD, semantic AD only focuses on the semantic shift, where covariate shift does not exist. An example of real-world applications is crime surveillance [43], [44]. Active image crawlers for a specific category also need semantic AD methods to ensure the purity of the collected images [45]. 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.

Even with different settings, many sensory and semantic AD methods are shown to be mutually inclusive in Section 3.

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 true normalities as positive and anomalies as negative, 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) [46]. Similarly, the precision and recall values can be used to compute metrics of F-scores and the area under the precision-recall curve (AUPR) [47]. 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.

2.2 Novelty Detection

Background The word "novel" generally refers to the unknown, new, and something interesting. Although the goal of novelty detection (ND) is similar to that of AD, there are three differences to note: **(1)** 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 [18]; **(2)** Novelty detection primarily focuses on semantic shift, which is also known as "novel class detection"; **(3)** Novelty detection removes the restriction that the ID samples should belong to only one class. One or multiple classes can appear during training.

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 shifts but no covariate shift. Notice that since the in-distribution can contain one or more classes, we distinguish two sub-tasks: one-class ND and multi-class ND. Since one-class ND and semantic AD have the same problem definition except for some nuances in motivation (ref. Section 2.2), their solution space is shared and will be presented in Section 3.

Application and Benchmark Real-world ND application includes video surveillance [43], [44], planetary exploration [48] and incremental learning [49], [50]. 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.

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

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 [51]. Some literature refers to this notorious overconfident behavior of the model as "arrogance", or "agnostophobia" [52]. 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.

Definition OSR requires the multi-class classifier to simultaneously: (1) accurately classify test samples that 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.*, $P(Y) \neq P'(Y)$, but no intentional covariate shift. 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 $P(Y)$. We will introduce the methodologies for OSR and multi-class ND together in Section 4.

Application and Benchmark OSR supports the robust deployment of real-world image classifiers in general, which can reject unknown samples in the open world [53], [54]. 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 [52], 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 can overconfidently classify samples from different semantic distributions, the field of out-of-distribution detection emerges, requiring models to reject label-shifted samples to guarantee reliability and safety.

Definition Out-of-distribution detection aims to detect test samples with non-overlapping labels *w.r.t* training data. Formally, test samples in the OOD detection setting come from the distribution with semantic shift from ID, *i.e.*, $P(Y) \neq P'(Y)$. The ID can contain a single class or multiple classes. When multiple classes exist in training, OOD detection should NOT harm the ID classification capability.

Position in Framework Out-of-distribution detection is a super-category that includes semantic AD, one-class ND, multi-class ND, and open-set recognition. In the multi-class setting, the problem can be canonical to OSR (Section 4)—accurately classify test samples from ID within the class space $\mathcal{Y}$, and reject OOD test samples with semantics outside the support of $\mathcal{Y}$. However, OOD detection encompasses a broader spectrum of learning tasks (*e.g.*, multi-label classification [55], reinforcement learning) and solution space (*e.g.*, density estimation and outlier exposure). Some approaches

relax the constraints imposed by OSR and achieve strong performance. Moreover, OOD detection also includes other subtopics of one-class novelty detection and semantic anomaly detection where a single class exists in the in-distribution.

Application and Benchmark The application of OOD detection usually falls into safety-critical situations such as autonomous driving [56], [57]. 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 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 [58], 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 (usually high-risk) tasks, such as a medical diagnosis model that trained by one hospital should detect scans under distributional shift, research on this topic remains a controversial task *w.r.t* OOD generalization tasks (*c.f*. Section 2.6). Detecting semantic shift has been the mainstream of OOD detection tasks.

2.5 Outlier Detection

Background According to *Wikipedia* [59], 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 [13], [14], [15]. 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 [60].

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 [61], [62], [63], outlier detection is also used in the real-world computer vision applications such as video surveillance [64] and dataset cleaning [65], [66], [67]. 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 [68], webly supervised learning [69], and open-set semi-supervised learning [70]. 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.

2.6 Discussion

In this subsection, we further contrast and summarize how the five sub-topics described above fit in our *generalized OOD detection* framework. As shown in Figure 1, semantic AD and one-class ND have identical problem formulation, despite the subtle difference in motivation. Multi-class ND and OSR both focus on semantic shift *w.r.t* a multi-class classification model. The only difference is that multi-class ND does not require ID classification whereas OSR does.

Despite the practical relevance of OSR, several restrictions remain such as no additional data is permitted during training and a required guarantee on theoretical open-risk bound. These restrictions exclude methods that focus more on effective performance improvement but may violate the constraints of OSR. On the other hand, OOD detection encompasses a broader spectrum of learning tasks and solution space (to be discussed in Section 5). Some approaches relax the constraints imposed by OSR and achieve strong performance.

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. Additionally, we briefly discuss five related topics below, and further clarify the scope of this survey.

Learning with Rejection can date back to early works on abstention [71], [72], which considered simple model families such as SVMs [73]. The phenomenon of neural networks' overconfidence in out-of-distribution data is first revealed by [74]. Despite methodologies differences, subsequent works developed on OOD detection and OSR share the underlying spirit of classification with rejection option.

Domain Adaptation (DA) [12] and Domain Generalization (DG) [75] 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 regardless of the covariate shift [76]. 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 Section 7.2. The difference between DA and DG is that while the former requires extra but few training samples from the target domain, the latter one does not.

Novelty Discovery [77] requires all observations are 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 [78] 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 expect 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 [79] aims to build a lifelong learning machine that can actively detect novel images [80], label them as new classes, and perform continuous learning. It can be viewed as the combination of novelty detection and incremental learning.

3 Anomaly Detection & One-Class Novelty Detection: Methodology

In this section, we review methodologies for AD and one-class ND. Most of the methods for sensory AD and semantic AD are shared, except for sensory AD focuses more on local information in the images and internal information of neural networks. Moreover, semantic AD and one-class ND have the same problem formulation (recall Section 2.2), therefore we review the methods for these three sub-tasks altogether.

Given the homogeneous in-distribution data, a straightforward approach is to estimate the in-distribution density and reject OOD test samples that deviate from the estimated distribution. We summarize density-based methods in Section 3.1. Alternative methods rely on the quality of image reconstruction to distinguish anomalous samples (Section 3.2), or directly learn a decision boundary between ID and OOD data (Section 3.3). We also review distance-based and meta-learning-based methods in Section 3.4 and Section 3.5. Lastly, we conclude with a discussion and present theoretical works in Section 3.6.

3.1 Density-based Methods

Density-based methods attempt to model the distribution of normal data (ID), with an operating assumption that anomalous test data has low likelihood whereas normal data has higher likelihood under the estimated density model.

3.1.1 Classic Density Estimation

a. Parametric Density Estimation Parametric density estimation assumes the ID density can be expressed through some pre-defined distributions [81]. One approach is to fits a multivariate Gaussian distribution on the training data and measures the Mahalanobis distance between the test sample and the expectation of training samples [82], [266]. Other works adopt more complex assumptions on in-distribution, such as mixed Gaussian distribution [83], [267], and Poisson distribution [84], *etc*.

b. Non-parametric Density Estimation Nonparametric density estimation solves a more practical scenario where a predefined distribution is unable to model the real distribution [85]. One can simply model the training distribution with histograms [268], [269], [270], [271]. Kernel density estimation (KDE) further uses the kernel function as a continuous replacement for the discrete histogram [86], [272], [273]. It flexibly takes parameters such as point weights and bandwidth to control the estimated distribution.

Discussion Although the classic density estimation methods obtain strong AD performance on wide ranges of tasks [274], [275], they are better suited for low-dimensional data. For high-dimensional data in computer vision tasks, these methods suffer from computational and scalability issues due to the curse of dimensionality [276]. To alleviate the problem, shallow methods implement feature engineering to reduce the dimensionality [277], [278].

3.1.2 Density Estimation with Deep Generative Models

In the context of deep learning, neural networks can produce features with high representation quality, which significantly enhance the performance of classic density estimation.

a. AE/VAE-based Models An autoencoder (AE) learns efficient representations of unlabeled data by reconstructing the input from the latent embedding [279]. Variational autoencoder (VAE) [280] encodes input images into latent vectors under the Gaussian distribution. The learned encoding can be considered as the lower-dimensional representation of the input. Classic density estimation methods can then be applied on top of these deep representations [87], [88], [89].

b. GAN-based Models Generative adversarial networks (GANs) consist of a generative network and a discriminative network, contesting with each other in a zero-sum game [281]. Typically, the generative network learns to map from a latent space to a data distribution of interest, whereas the discriminative network distinguishes candidates produced by the generator from the true data distribution. However, unlike the previous AE/VAE paradigm, the lack of an encoder makes it difficult for a GAN to directly find the corresponding embedding for a given image. To solve the problem, ADGAN [90] searches for a good representation in the latent space for a given sample. If such a representation is not found, the sample is deemed anomalous. However, this method can be computationally expensive.

c. Flow-based Models A normalizing flow describes the transformation of a probability density through a sequence of invertible mappings. By repeatedly applying the rule for change of variables, the initial density "flows" through the sequence of invertible mappings [91], [92]. Therefore, methods with the normalizing flow can directly estimate the likelihood of the input space. The flow-based methods are appraised by their elegant mathematical presentations, and are also shown to be sensitive to low-level features only. Flow-based methods can lead to significant computational costs since no dimensionality reduction is performed.

d. Representation Enhancement Apart from obtaining visual embeddings through generative models, some methods focus on enhancing the model capacity to increase

TABLE 1: Paper list for generalized out-of-distribution detection tasks.

Task	Methodology		Reference
§ 3 Anomaly Detection & One-Class Novelty Detection	§ 3.1 Density	§ 3.1.1: Classic Density Est.	[81], [82], [83], [84], [85], [86]
		§ 3.1.2: NN-based Density Est.	[87], [88], [89], [89], [90], [91], [92], [93], [94], [95], [96], [97], [98]
		§ 3.1.3: Energy-based Models	[99], [100], [101]
		§ 3.1.4: Frequency-based Methods	[102], [103], [104], [105]
	§ 3.2 Reconstruction	§ 3.2.1: Sparse Representation	[106], [107], [108], [109], [110]
		§ 3.2.2: Reconstruction-Error	[89], [111], [111], [112], [112], [113], [113], [114], [115], [116], [117], [118], [119], [120], [121], [122], [123], [124]
	§ 3.3 Classification	§ 3.3.1: One-Class Classification	[125], [126], [127], [128]
		§ 3.3.2: PU Learning	[129], [130], [131], [132], [133], [134], [135], [136], [137], [138], [139], [140], [141]
		§ 3.3.3: Self-Supervised Learning	[142], [143], [144], [145], [146]
	§ 3.4: Distance-based Methods		[147], [148], [149]
	§ 3.5: Gradient-based Methods		[150]
	§ 3.6: Discussion and Theoretical Analysis		[151], [152]
§ 4 Multi-Class Novelty Detection & Open Set Recognition	§ 4.1 Classification	§ 4.1.1: EVT-based Calibration	[153], [154], [155], [156]
		§ 4.1.2: EVT-free Calibration	[157], [158], [159]
		§ 4.1.3: Unknown Generation	[160], [161], [162], [163], [164], [165], [166]
		§ 4.1.4: Label Space Redesign	[167], [168], [169], [170], [171], [172]
	§ 4.2: Distance-based Methods		[173], [174], [175], [176], [177], [178], [179]
	§ 4.3 Reconstruction	§ 4.3.1: Sparse Representation	[180], [181], [182]
		§ 4.3.2: Reconstruction-Error	[183], [184], [185], [186]
§ 5 Out-of-Distribution Detection	§ 5.1 Classification	§ 5.1.1.a: Post-hoc Detection	[55], [187], [188], [189], [190], [191], [192]
		§ 5.1.1.b: Conf. Enhancement	[58], [192], [193], [194], [195], [196], [197], [198], [199], [200], [201], [202], [203], [204], [205], [206], [207], [208], [209]
		§ 5.1.1.c: Outlier Exposure (OE)	[52], [210], [211], [212], [213], [214], [215], [216], [217], [218], [219]
		§ 5.1.2: OOD Data Generation	[220], [221], [222], [223]
		§ 5.1.3: Gradient-based Methods	[188], [191]
		§ 5.1.4: Bayesian Models	[224], [225], [226], [227], [228], [229]
		§ 5.1.5: Large-scale OOD Detection	[168], [171], [230], [231]
	§ 5.2: Density-based Methods		[87], [88], [89], [90], [92], [121], [207], [232], [233], [234], [235], [236], [237], [238], [239], [240]
	§ 5.3: Distance-based Methods		[207], [241], [242], [243], [244], [245], [246]
§ 6 Outlier Detection	§ 6.1: Density-based Methods		[247], [248], [249], [250], [251], [252], [253], [254]
	§ 6.2 Distance	§ 6.2.1: Cluster-based Methods	[255], [256]
		§ 6.2.2: Graph-based Methods	[257], [258], [259], [260], [261]
	§ 6.3: Classification-based Methods		[125], [126], [142], [262], [263], [264], [265]

the representation power of the extracted features, which may better characterize the normality/ID-ness for more accurate density estimation. These strategies include data augmentation [93], adversarial training [89], distillations [94], loss function enhancement [95], and usage of shallow [96], [97] and local features [98].

3.1.3 *Energy-based Models*

Energy-based model (EBM) is a generative model that uses a scalar energy score to express the probability density of variables through unnormalized negative log probability [282], which provides a valid solution for AD [99]. However, compared to standard deep learning models, the training process of EBMs can be computationally expensive, since MCMC sampling and approximations are required to calculate integrals. To address the problem, methods such as the score matching method [100] and stochastic gradient Langevin dynamics [101] are proposed for efficient training.

3.1.4 *Frequency-based Methods*

Previous works also explored frequency domain analysis for anomaly detection. While humans perceive images based on low-frequency components, CNN models can largely depend on high-frequency components for decision making [102], [103]. Methods such as CNN kernel smoothing [102] and spectrum-oriented data augmentation [104] are proposed to

suppress the influence of high-frequency components. Other works also find that adversarial attacks on low-frequency components are also difficult to detect, therefore proposing methods to target the phase spectrum [105]. Frequency-based methods mainly focus on sensory AD (especially on detecting adversarial examples), which may not be suitable for semantic AD.

3.2 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. The difference of model performance can be measured in the feature space (Section 3.2.1) or by the reconstruction error (Section 3.2.2).

3.2.1 Sparse Representation

Sparse reconstruction assumes that every normal sample can be reconstructed accurately using a limited set of basis functions, whereas anomalous data should suffer from larger reconstruction costs, thus generating a dense representation [106], [107], [108]. Exemplar techniques for sparse encoding include L_1 norm-based kernel PCA [109] and low-rank embedded networks [110].

3.2.2 Reconstruction-Error Methods

Reconstruction-error methods rely on the assumption that a reconstruction model trained on the normal data will produce higher-quality outcomes for normal test samples as opposed to anomalies. Deep reconstruction models include AEs [111], VAEs [112], GANs [113], and U-Net [114] that can all be used as the backbone for this method.

a. AE/VAE-based Models Apart from the standard combination of reconstruction-error and AE/VAE models [111], [112], other methods use more sophisticated strategies such as reconstructing by memorized normality [115], [116], adapting model architectures [117], and partial/conditional reconstruction [89], [118], [119]. In the semi-supervised AD setting, CoRA [120] trains two AEs on inliers and outliers respectively. The reconstruction errors derived from the two AEs can be used as an indicator of anomaly.

b. GAN-based Models Advancement in generative modeling has led to the remarkable development of reconstruction-error methods using GANs. The discriminator in GANs intrinsically calculates the reconstruction error for anomaly detection [113]. Moreover, variants of GANs—such as denoising GANs [121] and class-conditional GANs [122]—enable further performance improvement by increasing the reconstruction difficulty. Some methods utilize the performance of the reconstructed image in downstream tasks to further amplify the reconstruction error of anomalies [123]. Ensembling can also enhance the performance [124].

3.3 Classification-based Methods

AD and one-class ND is often formulated as an unsupervised learning problem, where the entire ID data belongs to one class. Fully supervised AD is studied in [283]. The idea of classifier boundaries is successfully implemented and marked as a one-class classification task [125], [284], which we describe in Section 3.3.1. When it comes to semi-supervised AD setting where unlabeled data is introduced for training, PU learning is proposed for this specific problem, which will be introduced in Section 3.3.2. Lastly, we introduce self-supervised learning methods in Section 3.3.3.

3.3.1 One-Class Classification

One-class classification (OCC) directly learns a decision boundary that corresponds to a desired density level set of the normal data distribution [125]. DeepSVDD [126] first introduced the classic OCC to the deep learning community, which maps normal/ID examples into a hypersphere so that the description of normality is bounded. Deviations from this description are then deemed to be anomalies. Later, some works try to extend the method through elastic regularization [127] or constructing an adapted description with multi-linear hyperplanes [128].

3.3.2 Positive-Unlabeled Learning

Positive-unlabeled learning, or PU learning, focuses on the semi-supervised AD setting where unlabeled data is available in addition to the normal data [129], [130], [131]. The unlabeled data can contain both positive and negative examples. Popular PU learning methods generally rely on two strategies. One approach is to select reliable negative samples from unlabeled data and convert them into the supervised AD setting. Techniques such as distance to prototypes [132], [133], [134], clustering [135], [136], and density-based models [137] are used to filter out reliable negatives. Others consider the entire unlabeled set as noisy negatives, converting it into learning with noisy labels setting. Techniques such as sample re-weighting [138] and label cleaning methods [139], [140] have also shown their effectiveness for the task. Besides, reconstruction-error methods can be re-purposed for PU learning by training two reconstruction models for the positive and unlabeled set, respectively [141]. The comparison between their reconstruction-error scores indicates the final AD decision.

3.3.3 Self-Supervised Learning

Self-supervised learning methods tackle the AD and one-class ND problems in two aspects: (1) the enhancement of feature quality can improve AD performance; (2) some well-designed surrogate tasks can help reveal the anomalies from normal samples. In this part, we only discuss the second pretext task designing, since the first methods that improve feature quality are introduced with their corresponding main tasks, such as in Section 3.1.2. One classic method is isolation forest [142], which generates a random forest to contrast every normal sample. A test anomaly can be isolated in fewer steps than normal instances. Other methods use pretext tasks such as contrastive learning [143] and image transformation prediction [144], [145], where anomalies are more likely to make mistakes on the designed task. For video data, a natural self-supervised task is to predict future frames based on the existing ones [146], where larger error indicates abnormalities.

3.4 Distance-based Methods

Distance-based methods detect anomalies by calculating the distance between targeted samples and a number of internally stored exemplars, or prototypes [285]. These methods usually require training data in the memory. Representative methods include K-nearest Neighbors [147], prototype-based methods [148], [149], as well as methods to be introduced in Section 6.2.

3.5 Gradient-based Methods

Gradient-based method belongs to meta-learning or learning to learn, which is a topic of systematically observing the internal mechanisms of the learning tasks or models to propose methods based on the learned experience, or meta-data [286], [287]. To address AD tasks, some method observes the different patterns on training gradient between normalities and anomalies in a reconstruction task and hence use gradient-based representation to characterize anomalies [150].

3.6 Discussion

Sensory *vs.* Semantic AD Sensory and semantic AD both consider the normality as homogeneous, even though there might be multiple categories in the normal data. Solutions to semantic AD are mostly applicable to sensory AD problems. In particular, sensory AD problems can benefit from methods that focus on lower-level features (*e.g.*, flow-based and hidden feature-based), local representations, and frequency-based methods (*c.f*. Section 3.1.4).

Theoretical Analysis In addition to algorithmic development, several works provided theoretical analysis on AD and one-class ND. In [151], a clean set of ID and a mixed set of ID/OOD are constructed with identical sample sizes. A PAC-style finite sample guarantee is achieved for a certain probability of detecting a certain portion of anomalies with the minimum number of false alarms. Furthermore, in [152], a generalization error bound is provided for PU learning methods in semi-supervised AD.

Anomaly Detection *vs.* Outlier Detection If we model the test samples and training samples altogether, the AD problem will be transformed into an OD problem, and therefore the transductive approaches in Section 6 are also applicable. However, this method requires all training data to estimate test abnormality, which greatly increases the storage burden and computational complexity. Therefore, we do not include these methods in this part, but leave it to Section 6.

4 Open Set Recognition & Multi-Class Novelty Detection: Methodology

In this section, we introduce the methodology for multi-class ND and open-set recognition (OSR) together. We discuss these two sub-tasks together since both tasks focus on the scenario where ID data comprises multiple classes. The only difference is that OSR has an additional objective to accurately classify the ID data, while multi-class ND produces an ID/OOD binary classifier.

Since multi-class ND and OSR consider multiple classes during training, most methods are classification-based (Section 4.1). Alternative methods are based on ID prototypes (Section 4.2) and reconstruction (Section 4.3). Few density-based methods will be discussed in Section 4.4 along with a discussion.

4.1 Classification-based Methods

The concept of OSR was first introduced in [51], which showed the validity of 1-class SVM and binary SVM for solving the OSR problem. In particular, [51] 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.

4.1.1 EVT-based Uncertainty Calibration

Early works observe the overconfidence of neural networks and therefore focus on redistributing the logits by using the compact abating probability (CAP) [153] and extreme value theory (EVT) [154], [288], [289]. 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 [155] 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. OpenMax also provides alternative solutions by using penultimate features for EVT modeling, forming a density-based method.

4.1.2 EVT-Free Confidence Enhancement

To circumvent the requirement of constructing open-set risks, some works achieved good empirical results without EVT. For example, [157] 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. Apart from explicitly forcing discrepancy between known/unknown classes, other methods extract stronger features through an auxiliary task of transformation classification [158], or mutual information maximization between the input image and its latent features [159], *etc*.

4.1.3 Unknown Class Generation

Image generation techniques have been utilized to synthesize unknown samples from known classes, which helps distinguish between known vs. unknown samples [160], [161], [162], [163]. 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 [164], [165]. Moreover, [166] splits the training data into typical and atypical subsets, which also helps learn compact classification boundaries.

4.1.4 Label Space Redesign

Both OSR and multi-class ND focus on the settings when ID contains more than one category. One-hot encoding is commonly used to encode categorical information for classification. However, 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 novel classes detection. Some works arrange the large semantic space into a hierarchical taxonomy of known classes [167], [168]. Under the redesigned label architecture, top-down classification strategy [167] and group softmax training [168] are demonstrated effective. Another set of works uses word embeddings to automatically construct the label space. In [169], 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 [170] to better detect novel classes, where the image encoding space also contains rich information from the label space [171], [172].

4.2 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 [173], [174]. Some methods also leveraged contrastive learning to learn more compact clusters for known classes [175], [176], which enlarge the distance between ID and OOD. CROSR [177] 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 [178] 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 [179]. By storing the training samples, the nearest neighbor distance ratio is used for identifying unknown samples in testing.

4.3 Reconstruction-based Methods

With similar motivations as Section 3.2, 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.

4.3.1 Sparse Representation Methods

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 [180] and kernel null space methods [181], [182] are used for sparse encoding.

4.3.2 Reconstruction-Error Methods

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 [183]. 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 reject unknown samples [184]. Other methods generate counterfactual images, which help the model focus more on semantics [185]. Adversarial defense is also considered in [186] to enhance model robustness.

4.4 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 [155]. 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 [290].

Due to the restriction on using only ID data for training, OSR methods do not implement background classes, or outlier exposure (more in Section 5.1.1). We proceed by reviewing the recent OOD detection literature, which encompasses a broader problem space and methodological solutions to detecting the unknowns.

5 OOD Detection: Methodology

In this section, we introduce the methodology for OOD detection. We first present classification-based model in Section 5.1, followed by density-based methods in Section 5.2. Distance-based methods will be introduced in Sections 5.3. A brief discussion will be included at the end.

5.1 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 [187]. Early OOD detection methods focus on deriving improved OOD scores based on the output of neural networks.

5.1.1 Output-based Methods

a. Post-hoc Detection Post-hoc methods have the advantage of being easy to use without modifying the training 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 re-training can be prohibitive. Early work ODIN [188] 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

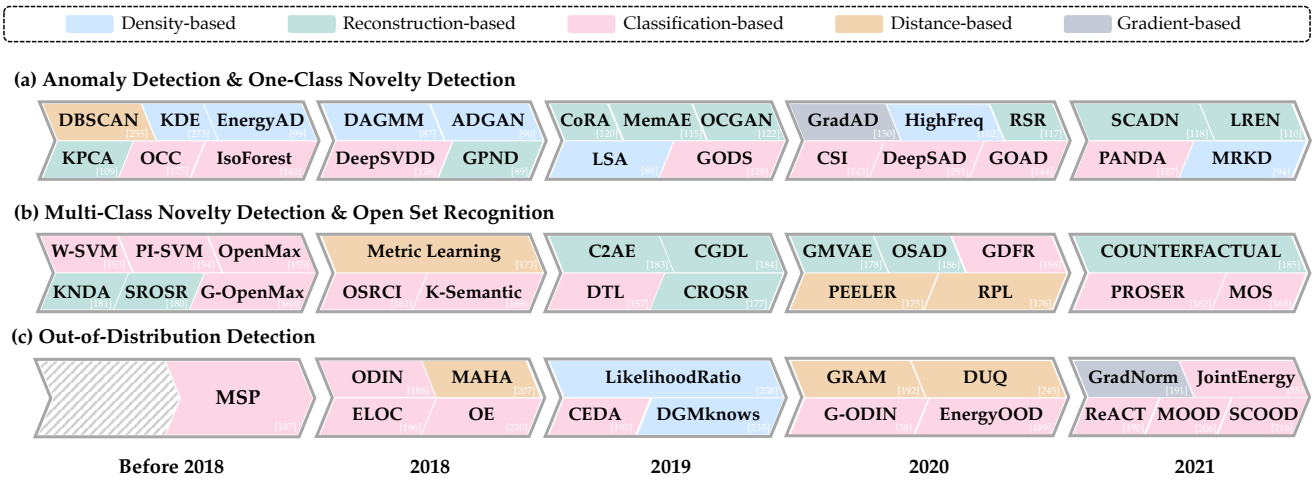


Fig. 3: Timeline for representative methodologies of **(a)** anomaly detection & one-class novelty detection, details in Section 3, **(b)** multi-class novelty detection & open set recognition, details in Section 4, and **(c)** OOD detection, details in Section 5, under generalized OOD detection framework. Different colors indicate different categories of methodologies. Each method has its corresponding reference (inconspicuous white) in the lower right corner. We do not list outlier detection methods in this figure due to their limited number of works on computer vision in deep learning era.

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 [189] proposed using an energy score for OOD detection, which is hyperparameter-free and achieves comparable or even better performance than ODIN. The energy function maps the logit outputs to a scalar through a convenient `logsumexp` operator. Test samples with lower energy are considered ID and vice versa. [55] further proposed JointEnergy score, which improves OOD uncertainty estimation for multi-label classification networks. Recently, [190] reveals one fundamental cause of the overconfidence issue on OOD data. In particular, 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. [190] provides a simple activation rectification strategy termed ReAct, which establishes strong post-hoc detection performance.

b. Confidence Enhancement Methods Tailored for OOD detection problem, confidence can be developed via designing a confidence-estimating branch [193] or class [194], data augmentation [195], ensembling with leaving-out strategy [196], adversarial training [197], [198], [199], [200], [214], stronger data augmentation [201], [202], [203], [204], certified certain uncertainty modeling [205], and utilizing feature from the optimal depth [206]. Specially, 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 [58] extended ODIN [188] by using a specialized training objective termed DeConf-C and choose 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. Techniques such as layer-wise Mahalanobis distance [207] and Gram Matrix [192], [208], [209] are implemented for better hidden feature quality to perform density estimation.

c. Outlier Exposure 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 [52], [210] and suppressing OOD feature magnitudes [52], a follow-up work, MCD [211] 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 [214], [217], [219]. A later work OECC [212] noticed that an extra regularization for confidence calibration introduces additional improvement for OE. To effectively utilize the given, usually massive, OOD samples, some works use outlier mining [214] and adversarial resampling [213] approaches to obtain a compact yet representative set. Other works consider a more practical scenario where given OOD samples contain ID samples, therefore using pseudo-labeling [215] or ID filtering methods [216] to reduce the interference of introduced ID. In general, OOD detection with outlier exposure can reach a much better performance. However, as research shows that the performance can be largely affected by the correlations between given and real OOD samples [218], further exploration is still required to generalize the exposed OOD samples to broader, universal OOD samples.

5.1.2 *OOD 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 [220], generate boundary samples

in the low-density area [221], or similarly, high-confidence OOD samples [222].

5.1.3 Gradient-based Methods

Existing OOD detection approaches primarily rely on the output (Section 5.1) or feature space for deriving OOD scores, while overlooking information from the gradient space. ODIN [188] 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 [191] 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.

5.1.4 Bayesian Models

A Bayesian model is a statistical model that implements Bayes' rule to infer all uncertainty within the model [291]. The most representative method is the Bayesian neural network [292], which draws samples from the posterior distribution of the model via MCMC [293], Laplace methods [294], [295] and variational inference [296], forming the epistemic uncertainty of the model prediction. However, their obvious shortcomings of inaccurate predictions [297] and high computational costs [298] prevent them from wide adoption in practice. Recent works attempt several less principled approximations including MC-dropout [224] and deep ensembles [225], [299] 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 [226]. 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 [227], [228], [229].

5.1.5 Large-scale OOD Detection

Recent works have advocated for OOD detection in large-scale settings, which are closer to real-world applications. Research efforts include scaling OOD detection to large semantic label space and exploiting large pre-trained models. For example, [168] revealed that approaches developed on the CIFAR benchmark might not translate effectively into ImageNet benchmark with a large semantic space, highlighting the need to evaluate OOD detection in a large-scale real-world setting. To overcome the challenge, the key idea of MOS [168] is to decompose the large semantic space into smaller groups with similar concepts, which allows simplifying the decision boundaries between known vs. unknown data. Recently, powerful pre-trained models have achieved astonishing results on various tasks and modalities. Several concurrent works [171], [230], [231] demonstrate that strong pretrained transformers can significantly improve some particularly difficult OOD tasks.

5.2 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 Section 3.1.2 can be directly adapted to OOD detection by unifying the ID data as a whole [87], [88], [89], [90], [121]. 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 [207]. Flow-based methods [92], [232], [233], [234] can also be used for probabilistic modeling. While directly estimating the likelihood seems like a natural approach, some works [235], [236], [237] find that probabilistic models sometimes assign a higher likelihood for the OOD sample. Several works attempt to solve the problems using likelihood ratio [238]. [239] finds that the likelihood exhibits a strong bias towards the input complexity and proposes a likelihood ratio-based method to compensate the influence of input complexity. Recent methods turn to new scores such as likelihood regret [240] or an ensemble of multiple density models [236]. Overall, generative models can be prohibitively challenging to train and optimize, and the performance can often lag behind the classification-based approaches (Section 3.3).

5.3 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. [207] 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 [241]. Some works use cosine similarity between test sample features and class features to determine OOD samples [242], [243]. 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 [244]. Moreover, other works leverage distances with radial basis function kernel [245] and Euclidean distance [246] between the input's embedding and the class centroids.

5.4 Discussion

The field of OOD detection has enjoyed rapid development since its emergence, with a large space of solutions ranging from classification-based to density-based to distance. In the multi-class setting, the problem can be canonical to OSR (Section 4)—accurately classify test samples from ID within the class space $\mathcal{Y}$, and reject OOD test samples with semantics outside the support of $\mathcal{Y}$. However, OOD detection encompasses a broader spectrum of learning tasks (*e.g.*, multi-label classification [55]) and solution space (*e.g.*, density estimation and outlier exposure). Some approaches relax the constraints imposed by OSR and achieve strong performance.

6 Outlier Detection: Methodology

Outlier detection (OD) requires the observation of all samples and aims to detect those that deviate significantly from the majority distribution. OD approaches are usually transductive, rather than inductive. Several surveys reviewed methodologies on this topic, yet mostly within the field of data mining [13], [14], [15], [16]. In this section, we briefly review OD methods, especially those developed for computer vision tasks using deep neural networks. We find that although deep learning methods rarely directly solve the OD problem, the data cleaning procedure, which is the prerequisite procedure of learning from open-set noisy data [68], [69] and open set semi-supervised learning [70], are solving the OD tasks.

6.1 Density-based Methods

A basic idea of OD models the entire dataset as a Gaussian distribution and flags samples that are at least three standard deviations from the mean [300], [301]. Other parametric probabilistic methods make use of Mahalanobis distance [266] and Gaussian mixtures [302] to model the data density. Similar to the "three standard deviations" rules, the interquartile range can also be used to identify outliers [247], forming a classic non-parametric probabilistic method. Local outlier factor (LOF) estimate the density of a given point via the ratio of the local reachability of its neighbors and itself [248], followed by derivations for robustification [249], [250] and simplification [251]. RANSAC [252] iteratively estimates parameters of a mathematical model to fit the data and find the samples as outliers that contribute less to estimates. Generally, classic density methods for AD such as kernel density estimation (*c.f*. Section 3.1) are also applicable for OD. Although these methods suffer from the curse of dimensionality on images, they can be alleviated by dimensionality reduction methods [253], [254] and the NN-based density methods (*c.f*. Section 3.1).

6.2 Distance-based Methods

A simple method to detect outliers is counting the number of neighbors within a certain radius, or measure the kth-nearest neighbor distance [303], [304]. We mainly discuss cluster-based methods and graph-based methods here.

6.2.1 Cluster-based Methods

DBSCAN [255] accumulates samples based on the distance-based density to form clusters. Samples that lie outside the major clusters are recognized as outliers. Subsequent works improve the clustering approaches by considering the confidence of cluster labels [256].

6.2.2 Graph-based Methods

Another set of methods uses the relationship among data points and constructs a neighborhood graph [305], [306] or its variants [307]. Graph properties and graph mining techniques are employed to find abnormal samples [257], [258], such as graph-based clustering [259], [260], partitioning [308], and label propagation with graph neural networks [261].

6.3 Classification-based Methods

AD methods (*e.g*., Isolation Forest [142], OC-SVM [125], [126], *etc*.) are also applicable to OD setting. When there are multiple classes in the dataset, researchers find that deep learning models—when trained with outliers—can still show robust prediction capability and identify the outliers [262]. Data cleaning using large pre-trained models is also common in the industry. Techniques to enhance model robustness and feature generalizability can be useful for this task, such as ensembling [263], co-training [264], and distillation [262], [265].

6.4 Discussion

Although the application of OD is not as common as other sub-tasks in the computer vision community, techniques for OD can be valuable for other tasks such as open-set semi-supervised learning, learning with open-set noisy labels, and potentially novelty discovery. In addition to methods reviewed here, most solutions to AD/ND/OOD detection can also be applied by considering all observations as ID (for model training) and then applying the model again on all the observations. In this case, methods such as reconstruction-based PCA (*c.f*. Section 3.2) and energy-based models (*c.f*. Section 3.1.3 and 5.1.1) can also address the OD problem.

7 Challenges and Future Directions

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

7.1 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 [172], [309], 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 [187] ignored the fact that some OOD datasets may contain images with ID categories, causing inaccurate performance evaluation. Fortunately, recent OOD detection works [168], [216] 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 [210] 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. As with the recent taken-down of TinyImages dataset [310], it poses a reproducibility crisis for OE-based methods. 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, [80] 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. [311] also showed that the ability of detecting novel classes can be highly correlated with its accuracy on the closed-set classes. [216] 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 have been primarily evaluated on small data sets such as CIFAR. It's been shown that approaches developed on the CIFAR benchmark might not translate effectively into ImageNet benchmark with a large semantic space, highlighting the need to evaluate OOD detection in a large-scale real-world setting. Therefore, we encourage future research to evaluate on ImageNet-based OOD detection benchmark [168], as well as large-scale OSR benchmark [311], and test the limits of the method developed. Moreover, real-world benchmarks that go beyond image classification can be valuable for the research community. In particular, for safety-critical settings such as autonomous driving and medical imaging diagnosis, more specialized benchmarks are needed and should be carefully constructed.

7.2 Future Directions

a. Methodologies across Sub-tasks Due to the inherent connections among different sub-tasks, their solution space can be shared and inspired from each other. For example, the recent emerging density-based OOD detection research (*c.f.* Section 5.2) can draw insights from the density-based AD methods (*c.f.* Section 3.1) 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 [216] that encourages detecting semantic OOD samples while being robust to negligible covariate shift. Given the vague definition of OOD, [312] proposed a new formalization of OOD detection by explicitly taking into account the separation between invariant features (semantic 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 label-shifted OOD data contains the spurious feature. Recent works on open long-tailed recognition [80], open compound domain adaptation [76], open-set domain adaptation [313] and open-set domain generalization [314] 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 [80], 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 [68], [315]. However, the open-set noisy samples can be useful for outlier exposure (*c.f.* 5.1.1) [308] 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 between OOD detection and the previous two fields can provide more insights and possibilities.

d. Theoretical Analysis While most of the existing OOD detection works focus on developing effective approaches to obtain better empirical performance, the theoretical analysis remains largely untapped. We hope future research can also contribute theoretical analyses and provide in-depth insights that help guide algorithmic development with rigorous guarantees.

8 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.

Acknowledgments

This study is supported by NTU NAP, and the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contributions 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).

REFERENCES

[1] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016. 1

[2] T. G. Dietterich, "Steps toward robust artificial intelligence," *AI Magazine*, 2017. 1

[3] N. A. Smuha, "The eu approach to ethics guidelines for trustworthy artificial intelligence," *Computer Law Review International*, 2019. 1

[4] B. Shneiderman, "Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered ai systems," *TiiS*, 2020. 1

[5] S. Mohseni, H. Wang, Z. Yu, C. Xiao, Z. Wang, and J. Yadawa, "Practical machine learning safety: A survey and primer," *arXiv preprint arXiv:2106.04823*, 2021. 1

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012. 1

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *ICCV*, 2015. 1

[8] N. Drummond and R. Shearer, "The open world assumption," in *eSI Workshop*, 2006. 1

[9] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *ICLR*, 2017. 1

[10] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan, "A theory of learning from different domains," *Machine learning*, 2010. 1

[11] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *ICCV*, 2017. 1

[12] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, 2018. 1, 5

[13] C. C. Aggarwal and P. S. Yu, "Outlier detection for high dimensional data," in *ACM SIGMOD*, 2001. 1, 5, 13

[14] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, 2004. 1, 5, 13

[15] I. Ben-Gal, "Outlier detection," in *Data mining and knowledge discovery handbook*, 2005. 1, 5, 13

[16] H. Wang, M. J. Bah, and M. Hammad, "Progress in outlier detection techniques: A survey," *Ieee Access*, 2019. 1, 13

[17] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller, "A unifying review of deep and shallow anomaly detection," *Proceedings of the IEEE*, 2021. 1, 3

[18] G. Pang, C. Shen, L. Cao, and A. v. d. Hengel, "Deep learning for anomaly detection: A review," *arXiv preprint arXiv:2007.02500*, 2020. 1, 4

[19] S. Bulusu, B. Kailkhura, B. Li, P. K. Varshney, and D. Song, "Anomalous example detection in deep learning: A survey," *IEEE Access*, 2020. 1

[20] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019. 1

[21] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, 2014. 1

[22] D. Miljković, "Review of novelty detection methods," in *MIPRO*, 2010. 1

[23] M. Markou and S. Singh, "Novelty detection: a review—part 1: statistical approaches," *Signal processing*, 2003. 1

[24] M. Markou and S. Singh, "Novelty detection: a review—part 2: neural network based approaches," *Signal processing*, 2003. 1

[25] T. E. Boult, S. Cruz, A. R. Dhamija, M. Gunther, J. Henrydoss, and W. J. Scheirer, "Learning and the unknown: Surveying steps toward open world recognition," in *AAAI*, 2019. 1

[26] C. Geng, S.-j. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *TPAMI*, 2020. 1

[27] A. Mahdavi and M. Carvalho, "A survey on open set recognition," *arXiv preprint arXiv:2109.00893*, 2021. 1

[28] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015. 3

[29] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *ICLR*, 2018. 3

[30] J. Quiñonero-Candela, M. Sugiyama, N. D. Lawrence, and A. Schwaighofer, *Dataset shift in machine learning*. Mit Press, 2009. 3

[31] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *CVPR*, 2016. 3

[32] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, 2018. 3

[33] K. Patel, H. Han, and A. K. Jain, "Secure face unlock: Spoof detection on smartphones," *IEEE transactions on information forensics and security*, 2016. 3

[34] D. Wen, H. Han, and A. K. Jain, "Face spoof detection with image distortion analysis," *IEEE Transactions on Information Forensics and Security*, 2015. 3

[35] K. A. Nixon, V. Aimale, and R. K. Rowe, "Spoof detection schemes," in *Handbook of biometrics*, 2008. 3

[36] G. Polatkan, S. Jafarpour, A. Brasoveanu, S. Hughes, and I. Daubechies, "Detection of forgery in paintings using supervised learning," in *ICIP*, 2009. 3

[37] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer, "The deepfake detection challenge (dfdc) preview dataset," *arXiv preprint arXiv:1910.08854*, 2019. 3

[38] L. Jiang, Z. Guo, W. Wu, Z. Liu, Z. Liu, C. C. Loy, S. Yang, Y. Xiong, W. Xia, B. Chen, P. Zhuang, S. Li, S. Chen, T. Yao, S. Ding, J. Li, F. Huang, L. Cao, R. Ji, C. Lu, and G. Tan, "DeeperForensics Challenge 2020 on real-world face forgery detection: Methods and results," *arXiv preprint arXiv:2102.09471*, 2021. 3

[39] P. Yang, D. Baracchi, R. Ni, Y. Zhao, F. Argenti, and A. Piva, "A survey of deep learning-based source image forensics," *Journal of Imaging*, 2020. 3

[40] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *CVPR*, 2019. 3

[41] W.-H. Chu and K. M. Kitani, "Neural batch sampling with reinforcement learning for semi-supervised anomaly detection," in *ECCV*, 2020. 3

[42] D. J. Atha and M. R. Jahanshahi, "Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection," *Structural Health Monitoring*, 2018. 3

[43] H. Idrees, M. Shah, and R. Surette, "Enhancing camera surveillance using computer vision: a research note," *Policing: An International Journal*, 2018. 3, 4

[44] C. P. Diehl and J. B. Hampshire, "Real-time object classification and novelty detection for collaborative video surveillance," in *IJCNN*, 2002. 3, 4

[45] L.-J. Li and L. Fei-Fei, "Optimol: automatic online picture collection via incremental model learning," *IJCV*, 2010. 3

[46] T. Fawcett, "An introduction to roc analysis," *Pattern recognition letters*, 2006. 3

[47] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," *JMLT*, 2020. 3

[48] H. R. Kerner, D. F. Wellington, K. L. Wagstaff, J. F. Bell, C. Kwan, and H. B. Amor, "Novelty detection for multispectral images with application to planetary exploration," in *AAAI*, 2019. 4

[49] H. Al-Behadili, A. Grumpe, and C. Wöhler, "Incremental learning and novelty detection of gestures in a multi-class system," in *AIMS*, 2015. 4

[50] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *ICML*, 2017. 4

[51] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *TPAMI*, 2013. 4, 9

[52] A. R. Dhamija, M. Günther, and T. E. Boult, "Reducing network agnostophobia," in *NeurIPS*, 2018. 4, 7, 11

[53] E. Sorio, A. Bartoli, G. Davanzo, and E. Medvet, "Open world classification of printed invoices," in *Proceedings of the 10th ACM symposium on Document engineering*, 2010. 4

[54] H. Xu, B. Liu, L. Shu, and P. Yu, "Open-world learning and application to product classification," in *WWW*, 2019. 4

[55] H. Wang, W. Liu, A. Bocchieri, and Y. Li, "Can multi-label classification networks know what they don't know?," *NeurIPS*, 2021. 4, 7, 11, 12

[56] X. Huang, D. Kroening, W. Ruan, J. Sharp, Y. Sun, E. Thamo, M. Wu, and X. Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability," *Computer Science Review*, 2020. 5

[57] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *CVPR*, 2012. 5

[58] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data," in *CVPR*, 2020. 5, 7, 11

[59] Wikipedia contributors, "Outlier from Wikipedia, the free encyclopedia," 2021. [Online; accessed 12 August 2021]. 5

[60] M. Bianchini, A. Belahcen, and F. Scarselli, "A comparative study of inductive and transductive learning with feedforward neural networks," in *Conference of the Italian Association for Artificial Intelligence*, 2016. 5

[61] I. Ben-Gal, "Outlier detection," in *Data mining and knowledge discovery handbook*, 2005. 5

[62] S. Basu and M. Meckesheimer, "Automatic outlier detection for time series: an application to sensor data," *Knowledge and Information Systems*, 2007. 5

[63] Y. Dou, W. Li, Z. Liu, Z. Dong, J. Luo, and S. Y. Philip, "Uncovering download fraud activities in mobile app markets," in *ASONAM*, 2019. 5

[64] T. Xiao, C. Zhang, and H. Zha, "Learning to detect anomalies in surveillance video," *IEEE Signal Processing Letters*, 2015. 5

[65] H. Liu, S. Shah, and W. Jiang, "On-line outlier detection and data cleaning," *Computers & chemical engineering*, 2004. 5

[66] A. Loureiro, L. Torgo, and C. Soares, "Outlier detection using clustering methods: a data cleaning application," in *Proceedings of KDNet Symposium on Knowledge-based Systems*, 2004. 5

[67] J. Van den Broeck, S. Argeseanu Cunningham, R. Eeckels, and K. Herbst, "Data cleaning: detecting, diagnosing, and editing data abnormalities," *PLoS medicine*, 2005. 5

[68] Y. Wang, W. Liu, X. Ma, J. Bailey, H. Zha, L. Song, and S.-T. Xia, "Iterative learning with open-set noisy labels," in *CVPR*, 2018. 5, 13, 14

[69] X. Chen and A. Gupta, "Webly supervised learning of convolutional networks," in *ICCV*, 2015. 5, 13

[70] K. Cao, M. Brbic, and J. Leskovec, "Open-world semi-supervised learning," *arXiv preprint arXiv:2102.03526*, 2021. 5, 13

[71] C. Chow, "On optimum recognition error and reject tradeoff," *IEEE Transactions on Information Theory*, 1970. 5

[72] G. Fumera and F. Roli, "Support vector machines with embedded reject option," in *International Workshop on Support Vector Machines*, 2002. 5

[73] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, 1995. 5

[74] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *CVPR*, 2015. 5

[75] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *arXiv preprint arXiv:2103.02503*, 2021. 5

[76] Z. Liu, Z. Miao, X. Pan, X. Zhan, D. Lin, S. X. Yu, and B. Gong, "Open compound domain adaptation," in *CVPR*, 2020. 5, 14

[77] K. Han, A. Vedaldi, and A. Zisserman, "Learning to discover novel visual categories via deep transfer clustering," in *CVPR*, 2019. 6

[78] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *TIST*, 2019. 6

[79] A. Bendale and T. Boult, "Towards open world recognition," in *CVPR*, 2015. 6

[80] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *CVPR*, 2019. 6, 14

[81] G. Danuser and M. Stricker, "Parametric model fitting: From inlier characterization to outlier detection," *TPAMI*, 1998. 6, 7

[82] C. Leys, O. Klein, Y. Dominicy, and C. Ley, "Detecting multivariate outliers: Use a robust variant of the mahalanobis distance," *Journal of Experimental Social Psychology*, 2018. 6, 7

[83] E. Eskin, "Anomaly detection over noisy data using learned probability distributions," in *ICML*, 2000. 6, 7

[84] M. Turcotte, J. Moore, N. Heard, and A. McPhall, "Poisson factorization for peer-based anomaly detection," in *IEEE Conference on Intelligence and Security Informatics (ISI)*, 2016. 6, 7

[85] A. J. Izenman, "Review papers: Recent developments in nonparametric density estimation," *Journal of the American Statistical Association*, 1991. 6, 7

[86] W. Hu, J. Gao, B. Li, O. Wu, J. Du, and S. Maybank, "Anomaly detection using local kernel density estimation and context-based regression," *TKDE*, 2018. 6, 7

[87] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *ICLR*, 2018. 6, 7, 12

[88] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *CVPR*, 2019. 6, 7, 12

[89] S. Pidhorskyi, R. Almohsen, D. A. Adjeroh, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," in *NeurIPS*, 2018. 6, 7, 8, 12

[90] L. Deecke, R. Vandermeulen, L. Ruff, S. Mandt, and M. Kloft, "Image anomaly detection with generative adversarial networks," in *ECML&KDD*, 2018. 6, 7, 12

[91] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in *ICML*, 2015. 6, 7

[92] I. Kobyzev, S. Prince, and M. Brubaker, "Normalizing flows: An introduction and review of current methods," *TPAMI*, 2020. 6, 7, 12

[93] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "Cutpaste: Self-supervised learning for anomaly detection and localization," in *CVPR*, 2021. 7

[94] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee, "Multiresolution knowledge distillation for anomaly detection," in *CVPR*, 2021. 7

[95] A. K. Menon and R. C. Williamson, "A loss framework for calibrated anomaly detection," in *NeurIPS*, 2018. 7

[96] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu, "Multi-attentional deepfake detection," in *CVPR*, 2021. 7

[97] P. Yang, J. Chen, C.-J. Hsieh, J.-L. Wang, and M. Jordan, "Ml-loo: Detecting adversarial examples with feature attribution," in *AAAI*, 2020. 7

[98] M. Du, S. Pentyala, Y. Li, and X. Hu, "Towards generalizable deepfake detection with locality-aware autoencoder," in *CIKM*, 2020. 7

[99] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," in *ICML*, 2016. 7

[100] A. Hyvärinen and P. Dayan, "Estimation of non-normalized statistical models by score matching.," 2005. 7

[101] M. Welling and Y. W. Teh, "Bayesian learning via stochastic gradient langevin dynamics," in *ICML*, 2011. 7

[102] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *CVPR*, 2020. 7

[103] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry, "Adversarial examples are not bugs, they are features," in *NeurIPS*, 2019. 7

[104] G. Chen, P. Peng, L. Ma, J. Li, L. Du, and Y. Tian, "Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain," *ICCV*, 2021. 7

[105] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu, "Spatial-phase shallow learning: rethinking face forgery detection in frequency domain," in *CVPR*, 2021. 7, 8

[106] A. Adler, M. Elad, Y. Hel-Or, and E. Rivlin, "Sparse coding with anomaly detection," *Journal of Signal Processing Systems*, 2015. 7, 8

[107] A. Li, Z. Miao, Y. Cen, and Y. Cen, "Anomaly detection using sparse reconstruction in crowded scenes," *Multimedia Tools and Applications*, 2017. 7, 8

[108] X. Mo, V. Monga, R. Bala, and Z. Fan, "Adaptive sparse representations for video anomaly detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2013. 7, 8

[109] Y. Xiao, H. Wang, W. Xu, and J. Zhou, "L1 norm based kpca for novelty detection," *Pattern Recognition*, 2013. 7, 8

[110] K. Jiang, W. Xie, J. Lei, T. Jiang, and Y. Li, "Lren: Low-rank embedded network for sample-free hyperspectral anomaly detection," in *AAAI*, 2021. 7, 8

[111] Z. Chen, C. K. Yeo, B. S. Lee, and C. T. Lau, "Autoencoder-based network anomaly detection," in *Wireless Telecommunications Symposium*, 2018. 7, 8

[112] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, 2015. 7, 8

[113] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient gan-based anomaly detection," in *ICLR-W*, 2018. 7, 8

[114] W. Liu, W. Luo, D. Lian, and S. Gao, "Future frame prediction for anomaly detection–a new baseline," in *CVPR*, 2018. 7, 8

[115] D. Gong, L. Liu, V. Le, B. Saha, M. R. Mansour, S. Venkatesh, and A. v. d. Hengel, "Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection," in *CVPR*, 2019. 7, 8

[116] H. Park, J. Noh, and B. Ham, "Learning memory-guided normality for anomaly detection," in *CVPR*, 2020. 7, 8

[117] C.-H. Lai, D. Zou, and G. Lerman, "Robust subspace recovery layer for unsupervised anomaly detection," *ICLR*, 2020. 7, 8

[118] X. Yan, H. Zhang, X. Xu, X. Hu, and P.-A. Heng, "Learning semantic context from normal samples for unsupervised anomaly detection," in *AAAI*, 2021. 7, 8

[119] D. T. Nguyen, Z. Lou, M. Klar, and T. Brox, "Anomaly detection with multiple-hypotheses predictions," in *ICML*, 2019. 7, 8

[120] K. Tian, S. Zhou, J. Fan, and J. Guan, "Learning competitive and discriminative reconstructions for anomaly detection," in *AAAI*, 2019. 7, 8

[121] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *CVPR*, 2018. 7, 8, 12

[122] P. Perera, R. Nallapati, and B. Xiang, "Ocgan: One-class novelty detection using gans with constrained latent representations," in *CVPR*, 2019. 7, 8

[123] K. Zhou, Y. Xiao, J. Yang, J. Cheng, W. Liu, W. Luo, Z. Gu, J. Liu, and S. Gao, "Encoding structure-texture relation with p-net for anomaly detection in retinal images," in *ECCV*, 2020. 7, 8

[124] X. Han, X. Chen, and L.-P. Liu, "Gan ensemble for anomaly detection," *arXiv preprint arXiv:2012.07988*, 2020. 7, 8

[125] D. M. J. Tax, "One-class classification: Concept learning in the absence of counter-examples.," 2002. 7, 8, 13

[126] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *ICML*, 2018. 7, 8, 13

[127] T. Reiss, N. Cohen, L. Bergman, and Y. Hoshen, "Panda: Adapting pretrained features for anomaly detection and segmentation," in *CVPR*, 2021. 7, 8

[128] J. Wang and A. Cherian, "Gods: Generalized one-class discriminative subspaces for anomaly detection," in *CVPR*, 2019. 7, 8

[129] B. Zhang and W. Zuo, "Learning from positive and unlabeled examples: A survey," in *International Symposiums on Information Processing*, 2008. 7, 8

[130] J. Bekker and J. Davis, "Learning from positive and unlabeled data: A survey," *Machine Learning*, 2020. 7, 8

[131] K. Jaskie and A. Spanias, "Positive and unlabeled learning algorithms and applications: A survey," in *International Conference on Information, Intelligence, Systems and Applications*, 2019. 7, 8

[132] C. Wang, C. Ding, R. F. Meraz, and S. R. Holbrook, "Psol: a positive sample only learning algorithm for finding non-coding rna genes," *Bioinformatics*, 2006. 7, 8

[133] X. Li and B. Liu, "Learning to classify texts using positive and unlabeled data," in *IJCAI*, 2003. 7, 8

[134] B. Zhang and W. Zuo, "Reliable negative extracting based on knn for learning from positive and unlabeled examples.," *Journal of Computers*, 2009. 7, 8

[135] S. Chaudhari and S. Shevade, "Learning from positive and unlabelled examples using maximum margin clustering," in *ICONIP*, 2012. 7, 8

[136] L. Liu and T. Peng, "Clustering-based method for positive and unlabeled text categorization enhanced by improved tfidf.," *Journal of Information Science and Engineering*, 2014. 7, 8

[137] F. He, T. Liu, G. I. Webb, and D. Tao, "Instance-dependent pu learning by bayesian optimal relabeling," *arXiv preprint arXiv:1808.02180*, 2018. 7, 8

[138] A. Menon, B. Van Rooyen, C. S. Ong, and B. Williamson, "Learning from corrupted binary labels via class-probability estimation," in *ICML*, 2015. 7, 8

[139] C. Scott, "A rate of convergence for mixture proportion estimation, with application to learning from noisy labels," in *Artificial Intelligence and Statistics*, 2015. 7, 8

[140] J.-X. Zhong, N. Li, W. Kong, S. Liu, T. H. Li, and G. Li, "Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection," in *CVPR*, 2019. 7, 8

[141] K. Tian, S. Zhou, J. Fan, and J. Guan, "Learning competitive and discriminative reconstructions for anomaly detection," in *AAAI*, 2019. 7, 8

[142] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *ICDM*, 2008. 7, 8, 13

[143] J. Tack, S. Mo, J. Jeong, and J. Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," in *NeurIPS*, 2020. 7, 8

[144] L. Bergman and Y. Hoshen, "Classification-based anomaly detection for general data," in *ICLR*, 2020. 7, 8

[145] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," in *NeurIPS*, 2018. 7, 8

[146] M.-I. Georgescu, A. Barbalau, R. T. Ionescu, F. S. Khan, M. Popescu, and M. Shah, "Anomaly detection in video via self-supervised and multi-task learning," in *CVPR*, 2021. 7, 8

[147] J. Tian, M. H. Azarian, and M. Pecht, "Anomaly detection using self-organizing maps-based k-nearest neighbor algorithm," in *PHM Society European Conference*, 2014. 7, 9

[148] G. Münz, S. Li, and G. Carle, "Traffic anomaly detection using k-means clustering," in *GI/ITG Workshop MMBnet*, 2007. 7, 9

[149] I. Syarif, A. Prugel-Bennett, and G. Wills, "Unsupervised clustering approach for network anomaly detection," in *International conference on networked digital technologies*, 2012. 7, 9

[150] G. Kwon, M. Prabhushankar, D. Temel, and G. AlRegib, "Backpropagated gradient representations for anomaly detection," in *ECCV*, 2020. 7, 9

[151] S. Liu, R. Garrepalli, T. Dietterich, A. Fern, and D. Hendrycks, "Open category detection with pac guarantees," in *ICML*, 2018. 7, 9

[152] Z. Fang, J. Lu, A. Liu, F. Liu, and G. Zhang, "Learning bounds for open-set learning," in *ICML*, 2021. 7, 9

[153] W. J. Scheirer, L. P. Jain, and T. E. Boult, "Probability models for open set recognition," *TPAMI*, 2014. 7, 9

[154] L. P. Jain, W. J. Scheirer, and T. E. Boult, "Multi-class open set recognition using probability of inclusion," in *ECCV*, 2014. 7, 9

[155] A. Bendale and T. E. Boult, "Towards open set deep networks," in *CVPR*, 2016. 7, 9, 10

[156] A. Rozsa, M. Günther, and T. E. Boult, "Adversarial robustness: Softmax versus openmax," in *BMVC*, 2017. 7

[157] P. Perera and V. M. Patel, "Deep transfer learning for multiple class novelty detection," in *CVPR*, 2019. 7, 9

[158] P. Perera, V. I. Morariu, R. Jain, V. Manjunatha, C. Wigington, V. Ordonez, and V. M. Patel, "Generative-discriminative feature representations for open-set recognition," in *CVPR*, 2020. 7, 9

[159] X. Sun, H. Ding, C. Zhang, G. Lin, and K.-V. Ling, "M2iosr: Maximal mutual information open set recognition," *arXiv preprint arXiv:2108.02373*, 2021. 7, 9

[160] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative openmax for multi-class open set classification," in *BMVC*, 2017. 7, 9

[161] L. Neal, M. Olson, X. Fern, W.-K. Wong, and F. Li, "Open set learning with counterfactual images," in *ECCV*, 2018. 7, 9

[162] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Learning placeholders for open-set recognition," in *CVPR*, 2021. 7, 9

[163] S. Kong and D. Ramanan, "Opengan: Open-set recognition via open data generation," in *ICCV*, 2021. 7, 9

[164] C. Geng and S. Chen, "Collective decision for open set recognition," *TKDE*, 2020. 7, 9

[165] J. Jang and C. O. Kim, "One-vs-rest network-based deep probability model for open set recognition," *arXiv preprint arXiv:2004.08067*, 2020. 7, 9

[166] P. Schlachter, Y. Liao, and B. Yang, "Open-set recognition using intra-class splitting," in *EUSIPCO*, 2019. 7, 9

[167] K. Lee, K. Lee, K. Min, Y. Zhang, J. Shin, and H. Lee, "Hierarchical novelty detection for visual object recognition," in *CVPR*, 2018. 7, 10

[168] R. Huang and Y. Li, "Mos: Towards scaling out-of-distribution detection for large semantic space," in *CVPR*, 2021. 7, 10, 12, 13, 14

[169] G. Shalev, Y. Adi, and J. Keshet, "Out-of-distribution detection using multiple semantic label representations," in *NeurIPS*, 2018. 7, 10

[170] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021. 7, 10

[171] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," *arXiv preprint arXiv:2106.03004*, 2021. 7, 10, 12

[172] W. Gan, "Language guided out-of-distribution detection," 2021. 7, 10, 13

[173] M. Masana, I. Ruiz, J. Serrat, J. van de Weijer, and A. M. Lopez, "Metric learning for novelty and anomaly detection," in *BMVC*, 2018. 7, 10

[174] Y. Shu, Y. Shi, Y. Wang, T. Huang, and Y. Tian, "p-odn: prototype-based open deep network for open set recognition," *Scientific reports*, 2020. 7, 10

[175] B. Liu, H. Kang, H. Li, G. Hua, and N. Vasconcelos, "Few-shot open-set recognition using meta-learning," in *CVPR*, 2020. 7, 10

[176] G. Chen, L. Qiao, Y. Shi, P. Peng, J. Li, T. Huang, S. Pu, and Y. Tian, "Learning open set network with discriminative reciprocal points," in *ECCV*, 2020. 7, 10

[177] R. Yoshihashi, W. Shao, R. Kawakami, S. You, M. Iida, and T. Naemura, "Classification-reconstruction learning for open-set recognition," in *CVPR*, 2019. 7, 10

[178] A. Cao, Y. Luo, and D. Klabjan, "Open-set recognition with gaussian mixture variational autoencoders," *AAAI*, 2020. 7, 10

[179] P. R. M. Júnior, R. M. De Souza, R. d. O. Werneck, B. V. Stein, D. V. Pazinato, W. R. de Almeida, O. A. Penatti, R. d. S. Torres, and A. Rocha, "Nearest neighbors distance ratio open-set classifier," *Machine Learning*, 2017. 7, 10

[180] H. Zhang and V. M. Patel, "Sparse representation-based open set recognition," *TPAMI*, 2016. 7, 10

[181] P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler, "Kernel null space methods for novelty detection," in *CVPR*, 2013. 7, 10

[182] J. Liu, Z. Lian, Y. Wang, and J. Xiao, "Incremental kernel null space discriminant analysis for novelty detection," in *CVPR*, 2017. 7, 10

[183] P. Oza and V. M. Patel, "C2ae: Class conditioned auto-encoder for open-set recognition," in *CVPR*, 2019. 7, 10

[184] X. Sun, Z. Yang, C. Zhang, K.-V. Ling, and G. Peng, "Conditional gaussian distribution learning for open set recognition," in *CVPR*, 2020. 7, 10

[185] Z. Yue, T. Wang, Q. Sun, X.-S. Hua, and H. Zhang, "Counterfactual zero-shot and open-set visual recognition," in *CVPR*, 2021. 7, 10

[186] R. Shao, P. Perera, P. C. Yuen, and V. M. Patel, "Open-set adversarial defense," in *ECCV*, 2020. 7, 10

[187] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in *ICLR*, 2017. 7, 10, 13

[188] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *ICLR*, 2018. 7, 10, 11, 12

[189] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *NeurIPS*, 2020. 7, 11

[190] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," in *NeurIPS*, 2021. 7, 11

[191] R. Huang, A. Geng, and Y. Li, "On the importance of gradients for detecting distributional shifts in the wild," in *NeurIPS*, 2021. 7, 12

[192] C. S. Sastry and S. Oore, "Detecting out-of-distribution examples with gram matrices," in *ICML*, 2020. 7, 11

[193] T. DeVries and G. W. Taylor, "Learning confidence for out-of-distribution detection in neural networks," *arXiv preprint arXiv:1802.04865*, 2018. 7, 11

[194] Y. Wang, B. Li, T. Che, K. Zhou, Z. Liu, and D. Li, "Energy-based open-world uncertainty modeling for confidence calibration," in *ICCV*, 2021. 7, 11

[195] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem," in *CVPR*, 2019. 7, 11

[196] A. Vyas, N. Jammalamadaka, X. Zhu, D. Das, B. Kaul, and T. L. Willke, "Out-of-distribution detection using an ensemble of self supervised leave-out classifiers," in *ECCV*, 2018. 7, 11

[197] J. Bitterwolf, A. Meinke, and M. Hein, "Certifiably adversarially robust detection of out-of-distribution data," in *NeurIPS*, 2020. 7, 11

[198] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha, "Robust out-of-distribution detection for neural networks," *arXiv preprint arXiv:2003.09711*, 2020. 7, 11

[199] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem," in *CVPR*, 2019. 7, 11

[200] S. Choi and S.-Y. Chung, "Novelty detection via blurring," in *ICLR*, 2020. 7, 11

[201] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in *NeurIPS*, 2019. 7, 11

[202] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *CVPR*, 2019. 7, 11

[203] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *arXiv preprint arXiv:1708.04552*, 2017. 7, 11

[204] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "Augmix: A simple data processing method to improve robustness and uncertainty," *arXiv preprint arXiv:1912.02781*, 2019. 7, 11

[205] A. Meinke and M. Hein, "Towards neural networks that provably know when they don't know," *arXiv preprint arXiv:1909.12180*, 2019. 7, 11

[206] Z. Lin, S. D. Roy, and Y. Li, "Mood: Multi-level out-of-distribution detection," in *CVPR*, 2021. 7, 11

[207] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in *NeurIPS*, 2018. 7, 11, 12

[208] C. S. Sastry and S. Oore, "Detecting out-of-distribution examples with in-distribution examples and gram matrices," in *NeurIPS-W*, 2019. 7, 11

[209] A. G. Pacheco, C. S. Sastry, T. Trappenberg, S. Oore, and R. A. Krohling, "On out-of-distribution detection algorithms with deep neural skin cancer classifiers," in *CVPR-W*, 2020. 7, 11

[210] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *ICLR*, 2019. 7, 11, 13

[211] Q. Yu and K. Aizawa, "Unsupervised out-of-distribution detection by maximum classifier discrepancy," in *ICCV*, 2019. 7, 11

[212] A.-A. Papadopoulos, M. R. Rajati, N. Shaikh, and J. Wang, "Outlier exposure with confidence control for out-of-distribution detection," *Neurocomputing*, 2021. 7, 11

[213] Y. Li and N. Vasconcelos, "Background data resampling for outlier-aware classification," in *CVPR*, 2020. 7, 11

[214] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha, "Atom: Robustifying out-of-distribution detection using outlier mining," *ECML&PKDD*, 2021. 7, 11

[215] S. Mohseni, M. Pitale, J. Yadawa, and Z. Wang, "Self-supervised learning for generalizable out-of-distribution detection," in *AAAI*, 2020. 7, 11

[216] J. Yang, H. Wang, L. Feng, X. Yan, H. Zheng, W. Zhang, and Z. Liu, "Semantically coherent out-of-distribution detection," in *ICCV*, 2021. 7, 11, 13, 14

[217] S. Thulasidasan, S. Thapa, S. Dhaubhadel, G. Chennupati, T. Bhattacharya, and J. Bilmes, "An effective baseline for robustness to distributional shift," *arXiv preprint arXiv:2105.07107*, 2021. 7, 11

[218] A. Shafaei, M. Schmidt, and J. J. Little, "A less biased evaluation of out-of-distribution sample detectors," in *BMVC*, 2019. 7, 11

[219] S. Mohseni, M. Pitale, J. Yadawa, and Z. Wang, "Self-supervised learning for generalizable out-of-distribution detection," in *AAAI*, 2020. 7, 11

[220] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," 2018. 7, 11

[221] S. Vernekar, A. Gaurav, V. Abdelzad, T. Denouden, R. Salay, and K. Czarnecki, "Out-of-distribution detection in classifiers via generation," in *NeurIPS-W*, 2019. 7, 12

[222] K. Sricharan and A. Srivastava, "Building robust classifiers through generation of confident out of distribution examples," in *NeurIPS-W*, 2018. 7, 12

[223] D. Mandal, S. Narayan, S. K. Dwivedi, V. Gupta, S. Ahmed, F. S. Khan, and L. Shao, "Out-of-distribution detection for generalized zero-shot action recognition," in *CVPR*, 2019. 7

[224] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016. 7, 12

[225] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *NeurIPS*, 2017. 7, 12

[226] K. Osawa, S. Swaroop, A. Jain, R. Eschenhagen, R. E. Turner, R. Yokota, and M. E. Khan, "Practical deep learning with bayesian principles," in *NeurIPS*, 2019. 7, 12

[227] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in *NeurIPS*, 2018. 7, 12

[228] A. Malinin and M. Gales, "Reverse kl-divergence training of prior networks: Improved uncertainty and adversarial robustness," in *NeurIPS*, 2019. 7, 12

[229] J. Nandy, W. Hsu, and M. L. Lee, "Towards maximizing the representation gap between in-domain & out-of-distribution examples," in *NeurIPS*, 2020. 7, 12

[230] D. Hendrycks, X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song, "Pretrained transformers improve out-of-distribution robustness," *arXiv preprint arXiv:2004.06100*, 2020. 7, 12

[231] R. Koner, P. Sinhamahapatra, K. Roscher, S. Günnemann, and V. Tresp, "Oodformer: Out-of-distribution detection transformer," *arXiv preprint arXiv:2107.08976*, 2021. 7, 12

[232] E. Zisselman and A. Tamar, "Deep residual flow for out of distribution detection," in *CVPR*, 2020. 7, 12

[233] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," *NeurIPS*, 2018. 7, 12

[234] A. Van Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *ICML*, 2016. 7, 12

[235] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan, "Do deep generative models know what they don't know?," in *NeurIPS*, 2018. 7, 12

[236] H. Choi, E. Jang, and A. A. Alemi, "Waic, but why? generative ensembles for robust anomaly detection," *arXiv preprint arXiv:1810.01392*, 2018. 7, 12

[237] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Why normalizing flows fail to detect out-of-distribution data," in *NeurIPS*, 2020. 7, 12

[238] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. A. DePristo, J. V. Dillon, and B. Lakshminarayanan, "Likelihood ratios for out-of-distribution detection," in *NeurIPS*, 2019. 7, 12

[239] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque, "Input complexity and out-of-distribution detection with likelihood-based generative models," 2020. 7, 12

[240] Z. Xiao, Q. Yan, and Y. Amit, "Likelihood regret: An out-of-distribution detection score for variational auto-encoder," in *NeurIPS*, 2020. 7, 12

[241] J. Ren, S. Fort, J. Liu, A. G. Roy, S. Padhy, and B. Lakshminarayanan, "A simple fix to mahalanobis distance for improving near-ood detection," *arXiv preprint arXiv:2106.09022*, 2021. 7, 12

[242] E. Techapanurak, M. Suganuma, and T. Okatani, "Hyperparameter-free out-of-distribution detection using cosine similarity," in *ACCV*, 2020. 7, 12

[243] X. Chen, X. Lan, F. Sun, and N. Zheng, "A boundary based out-of-distribution classifier for generalized zero-shot learning," in *ECCV*, 2020. 7, 12

[244] A. Zaeemzadeh, N. Bisagno, Z. Sambugaro, N. Conci, N. Rahnavard, and M. Shah, "Out-of-distribution detection using union of 1-dimensional subspaces," in *CVPR*, 2021. 7, 12

[245] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *ICML*, 2020. 7, 12

[246] H. Huang, Z. Li, L. Wang, S. Chen, B. Dong, and X. Zhou, "Feature space singularity for out-of-distribution detection," *arXiv preprint arXiv:2011.14654*, 2020. 7, 12

[247] X. Wan, W. Wang, J. Liu, and T. Tong, "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range," *BMC medical research methodology*, 2014. 7, 13

[248] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *SIGMOD*, 2000. 7, 13

[249] J. Tang, Z. Chen, A. W.-C. Fu, and D. W. Cheung, "Enhancing effectiveness of outlier detections for low density patterns," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2002. 7, 13

[250] H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek, "Loop: local outlier probabilities," in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009. 7, 13

[251] E. Schubert, A. Zimek, and H.-P. Kriegel, "Local outlier detection reconsidered: a generalized view on locality with applications to spatial, video, and network outlier detection," *Data mining and knowledge discovery*, 2014. 7, 13

[252] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, 1981. 7, 13

[253] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 2011. 7, 13

[254] V. Sharan, P. Gopalan, and U. Wieder, "Efficient anomaly detection via matrix sketching," *NeurIPS*, 2018. 7, 13

[255] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *KDD*, 1996. 7, 13

[256] U. Rebbapragada and C. E. Brodley, "Class noise mitigation through instance weighting," in *ECML*, 2007. 7, 13

[257] L. Akoglu, H. Tong, and D. Koutra, "Graph based anomaly detection and description: a survey," *Data mining and knowledge discovery*, 2015. 7, 13

[258] C. C. Noble and D. J. Cook, "Graph-based anomaly detection," in *SIGKDD*, 2003. 7, 13

[259] Y. Kou, C.-T. Lu, and R. F. Dos Santos, "Spatial outlier detection: a graph-based approach," in *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2007. 7, 13

[260] Z. Mingqiang, H. Hui, and W. Qian, "A graph-based clustering algorithm for anomaly intrusion detection," in *International Conference on Computer Science & Education (ICCSE)*, 2012. 7, 13

[261] J. Yang, W. Chen, L. Feng, X. Yan, H. Zheng, and W. Zhang, "Webly supervised image classification with metadata: Automatic noisy label correction via visual-semantic graph," in *ACM Multimedia*, 2020. 7, 13

[262] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *CVPR*, 2017. 7, 13

[263] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox, "Self: Learning to filter noisy labels with self-ensembling," in *ICLR*, 2020. 7, 13

[264] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *NIPS*, 2018. 7, 13

[265] J. Yang, L. Feng, W. Chen, X. Yan, H. Zheng, P. Luo, and W. Zhang, "Webly supervised image classification with self-contained confidence," in *ECCV*, 2020. 7, 13

[266] R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart, "The mahalanobis distance," *Chemometrics and intelligent laboratory systems*, 2000. 6, 13

[267] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the em algorithm," *SIAM review*, 1984. 6

[268] J. Van Ryzin, "A histogram method of density estimation," *Communications in Statistics-Theory and Methods*, 1973. 6

[269] M. Xie, J. Hu, and B. Tian, "Histogram-based online anomaly detection in hierarchical wireless sensor networks," in *ICTSPCC*, 2012. 6

[270] A. Kind, M. P. Stoecklin, and X. Dimitropoulos, "Histogram-based traffic anomaly detection," *IEEE Transactions on Network and Service Management*, 2009. 6

[271] M. Goldstein and A. Dengel, "Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm," *KI-2012: Poster and Demo Track*, 2012. 6

[272] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, 1962. 6

[273] M. Desforges, P. Jacob, and J. Cooper, "Applications of probability density estimation to the detection of abnormal conditions in engineering," *Proceedings of the institution of mechanical engineers*, 1998. 6

[274] K. Nishiya, J. Hasegawa, and T. Koike, "Dynamic state estimation including anomaly detection and identification for power systems," in *IEE proceedings C (generation, transmission and distribution)*, 1982. 6

[275] P. Helman and G. Liepins, "Statistical foundations of audit trail analysis for the detection of computer misuse," *IEEE Transactions on software engineering*, 1993. 6

[276] P. D. Talagala, R. J. Hyndman, and K. Smith-Miles, "Anomaly detection in high-dimensional data," *Journal of Computational and Graphical Statistics*, 2020. 6

[277] A. Patcha and J.-M. Park, "An overview of anomaly detection techniques: Existing solutions and latest technological trends," *Computer networks*, 2007. 6

[278] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *CVPR*, 2010. 6

[279] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE journal*, 1991. 6

[280] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. 6

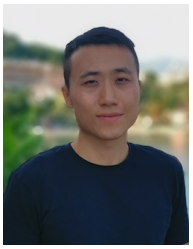
[281] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NIPS*, 2014. 6

[282] J. Ngiam, Z. Chen, P. W. Koh, and A. Y. Ng, "Learning deep energy models," in *ICML*, 2011. 7

[283] N. Görnitz, M. Kloft, K. Rieck, and U. Brefeld, "Toward supervised anomaly detection," *Journal of Artificial Intelligence Research*, 2013. 8

[284] S. S. Khan and M. G. Madden, "A survey of recent trends in one class classification," in *Irish conference on artificial intelligence and cognitive science*, 2009. 8
[285] D. Wettschereck, "A study of distance-based machine learning algorithms," 1994. 9
[286] J. Vanschoren, "Meta-learning: A survey," *arXiv preprint arXiv:1810.03548*, 2018. 9
[287] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *arXiv preprint arXiv:2004.05439*, 2020. 9
[288] R. L. Smith, "Extreme value theory," *Handbook of applicable mathematics*, 1990. 9
[289] E. Castillo, *Extreme value theory in engineering*. Elsevier, 2012. 9
[290] H. Zhang, A. Li, J. Guo, and Y. Guo, "Hybrid models for open set recognition," in *ECCV*, 2020. 10
[291] E. T. Jaynes, "Bayesian methods: General background," 1986. 12
[292] R. M. Neal, *Bayesian learning for neural networks*. 2012. 12
[293] D. Gamerman and H. F. Lopes, *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. CRC Press, 2006. 12
[294] D. J. C. Mackay, *Bayesian methods for adaptive models*. PhD thesis, California Institute of Technology, 1992. 12
[295] A. Y. Foong, Y. Li, J. M. Hernández-Lobato, and R. E. Turner, "'in-between' uncertainty in bayesian neural networks," in *ICML-W*, 2020. 12
[296] C. Peterson and E. Hartman, "Explorations of the mean field theory learning algorithm," *Neural Networks*, 1989. 12
[297] F. Wenzel, K. Roth, B. S. Veeling, J. Świątkowski, L. Tran, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, and S. Nowozin, "How good is the bayes posterior in deep neural networks really?," in *ICML*, 2020. 12
[298] A. Gelman, "Objections to bayesian statistics," *Bayesian Analysis*, 2008. 12
[299] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*, 2000. 12
[300] D. G. Altman and J. M. Bland, "Standard deviations and standard errors," *BMJ*, 2005. 13
[301] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of experimental social psychology*, 2013. 13
[302] X. Yang, L. J. Latecki, and D. Pokrajac, "Outlier detection with globally optimal exemplar-based gmm," in *SIAM*, 2009. 13
[303] M. Sugiyama and K. Borgwardt, "Rapid distance-based outlier detection via sampling," *NIPS*, 2013. 13
[304] G. H. Orair, C. H. Teixeira, W. Meira Jr, Y. Wang, and S. Parthasarathy, "Distance-based outlier detection: consolidation and renewed bearing," *Proceedings of the VLDB Endowment*, 2010. 13
[305] V. Hautamaki, I. Karkkainen, and P. Franti, "Outlier detection using k-nearest neighbour graph," in *ICPR*, 2004. 13
[306] F. Muhlenbach, S. Lallich, and D. A. Zighed, "Identifying and handling mislabelled instances," *Journal of Intelligent Information Systems*, 2004. 13
[307] W. Liu, J. He, and S.-F. Chang, "Large graph construction for scalable semi-supervised learning," in *ICML*, 2010. 13
[308] Z.-F. Wu, T. Wei, J. Jiang, C. Mao, M. Tang, and Y.-F. Li, "Ngc: A unified framework for learning with open-world noisy data," in *ICCV*, 2021. 13, 14
[309] F. Ahmed and A. Courville, "Detecting semantic anomalies," in *AAAI*, 2020. 13
[310] A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for nonparametric object and scene recognition," *TPAMI*, 2008. 14
[311] S. Vaze, K. Han, A. Vedaldi, and A. Zisserman, "Open-set recognition: A good closed-set classifier is all you need," *arXiv preprint arXiv:2110.06207*, 2021. 14
[312] Y. Ming, H. Yin, and Y. Li, "On the impact of spurious correlation for out-of-distribution detection," *arXiv preprint arXiv:2109.05642*, 2021. 14
[313] P. Panareda Busto and J. Gall, "Open set domain adaptation," in *ICCV*, 2017. 14
[314] Y. Shu, Z. Cao, C. Wang, J. Wang, and M. Long, "Open domain generalization with domain-augmented meta-learning," in *CVPR*, 2021. 14
[315] J. Li, C. Xiong, and S. C. Hoi, "Mopro: Webly supervised learning with momentum prototypes," *ICLR*, 2021. 14

Jingkang Yang is a Ph.D. student from the School of Computer Science and Engineering (SCSE), Nanyang Technological University (NTU) since Jan. 2021. He is affiliated with MMLab@NTU and S-Lab, supervised by Dr. Ziwei Liu. Previously, he received his Bachelor's degree in Telecommunications from Beijing University of Posts and Telecommunications and the Queen Mary University of London. His research topic is knowledge-integrated visual reasoning in the open world.



Kaiyang Zhou is currently a postdoctoral researcher at Nanyang Technological University, Singapore. He received his Ph.D. in 2020 from the University of Surrey, UK. His research centers around the development of robust, generalizable, and data-efficient learning algorithms and their applications in computer vision. His work has been published in top-tier venues, such as TPAMI, ICCV, and ICLR.

Yixuan (Sharon) Li is currently an Assistant Professor in the Department of Computer Sciences at the University of Wisconsin Madison. Previously she worked as a postdoc researcher in the Computer Science department at Stanford AI Lab (SAIL). She completed her Ph.D. from Cornell University in 2017, advised by John E. Hopcroft. She currently serves or has served as the Program Chair and founding organizer of the ICML Workshop on Uncertainty and Robustness in Deep Learning (UDL) during 2019-2021, Area Chair for NeurIPS, ICML, ICLR, AAAI, and IJCAI. Her works explore, understand, and mitigate the many challenges where failure modes can naturally occur in deploying machine learning models in the open world. She was named Forbes 30 Under 30 in Science, 30 Under 30 Rising Stars in AI, and JP Morgan early-career outstanding faculty.

Ziwei Liu is currently an Assistant Professor at Nanyang Technological University (NTU). Previously, he was a senior research fellow at the Chinese University of Hong Kong and a postdoctoral researcher at the University of California, Berkeley. Ziwei received his Ph.D. from the Chinese University of Hong Kong in 2017. His research revolves around computer vision/graphics, machine learning, and robotics. He has published extensively on top-tier conferences and journals in relevant fields, including CVPR, ICCV, ECCV, NeurIPS, IROS, SIGGRAPH, TOG, and TPAMI. He is the recipient of the Microsoft Young Fellowship, Hong Kong PhD Fellowship, ICCV Young Researcher Award, and HKSTP best paper award.