Recent Advances in OOD Detection: Problems and Approaches

SHUO LU, NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences, China
YINGSHENG WANG*, Anhui University, China
LIJUN SHENG, University of Science and Technology of China, China
AIHUA ZHENG, Anhui University, China
LINGXIAO HE, Meituan, China
JIAN LIANG, NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences, China

Out-of-distribution (OOD) detection aims to detect test samples outside the training category space, which is an essential component in building reliable machine learning systems. Existing reviews on OOD detection primarily focus on method taxonomy, surveying the field by categorizing various approaches. However, many recent works concentrate on non-traditional OOD detection scenarios, such as test-time adaptation, multi-modal data sources and other novel contexts. In this survey, we uniquely review recent advances in OOD detection from the problem scenario perspective for the first time. According to whether the training process is completely controlled, we divide OOD detection methods into training-driven and training-agnostic. Besides, considering the rapid development of pre-trained models, large pre-trained model-based OOD detection is also regarded as an important category and discussed separately. Furthermore, we provide a discussion of the evaluation scenarios, a variety of applications, and several future research directions. We believe this survey with new taxonomy will benefit the proposal of new methods and the expansion of more practical scenarios. A curated list of related papers is provided in the Github repository: https://github.com/shuolucs/Awesome-Out-Of-Distribution-Detection

CCS Concepts: • **Trustworthy Machine Learning → Out-of-distribution Detection**.

Additional Key Words and Phrases: Trustworthy Machine Learning, Out-of-distribution Detection

ACM Reference Format:
Shuo Lu, Yingsheng Wang, Lijun Sheng, Aihua Zheng, Lingxiao He, and Jian Liang. 2024. Recent Advances in OOD Detection: Problems and Approaches. 1, 1 (September 2024), 33 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Machine learning methods [1] have made significant progress under the closed-world assumption, where test data is drawn from the same distribution as the training set, known as in-distribution (ID). However, in the real world, models inevitably encounter test samples that do not belong to any training set category, commonly referred to as out-of-distribution (OOD) data. OOD detection [2] aims to identify and reject OOD samples rather than make overconfident predictions arbitrarily [3] while maintaining accurate classification for ID data. Models with superior OOD detection

*The first two authors contributed equally to this research.

Authors' addresses: Shuo Lu, NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China; Yingsheng Wang, Anhui University, Hefei, China; Lijun Sheng, University of Science and Technology of China, Hefei, China; Aihua Zheng, Anhui University, Hefei, China; Lingxiao He, Meituan, Beijing, China; Jian Liang, NLPR & MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing, China, liangjian92@gmail.com.

Manuscript submitted to ACM

capabilities are more reliable and have important applications in numerous security-critical scenarios. For instance, in medical diagnosis systems, a model that cannot detect OOD samples will misjudge unknown diseases and cause serious misdiagnosis [4]. Similarly, autonomous driving algorithms [5] should detect unknown scenarios and resort to human control to avoid accidents caused by arbitrary judgment.

Notably, several previous efforts have been dedicated to surveying and summarizing OOD detection in recent years. Yang et al. [6] discuss OOD detection with several similar topics and categorize existing work into classification-based, density-based, distance-based, and reconstruction-based methods. Cui and Wang [7] conduct a survey on OOD detection from a methodological perspective but with an alternative classification criterion, including supervised, semi-supervised, and unsupervised methods. Additionally, Lang et al. [8] offer a review of OOD detection methods in natural language processing. However, previous works focus too much on the discussion from the perspective of methods and lack an in-depth exploration from the viewpoint of task scenarios. Establishing a clear taxonomy of task scenarios can enhance a comprehensive understanding of the field and assist practitioners in selecting the appropriate method. Moreover, given the recent introduction of new paradigms (e.g., test-time learning paradigm [9]) and methods based on large pre-trained models [10], there is an urgent need for a survey that incorporates the latest technologies.

In this survey, we for the first time review the recent advances in OOD detection with a problem-oriented taxonomy, as illustrated in Fig. 1. Based on whether the method needs to control the pre-training process, we categorize OOD detection algorithms into *training-driven* and *training-agnostic* methods. Considering the rapid development of large pre-trained models nowadays, we also regard *large pre-trained model-based* OOD detection as a separate section. In detail, training-driven methods achieve high detection capability by designing the optimization process of the training stage. They are further classified and discussed according to whether OOD data is used in training. Training-agnostic methods distinguish OOD data from ID ones based on a well-trained model, skipping the time-consuming and expensive pre-training process in practice. According to whether utilizing test samples to further improve OOD detection performance, we categorize them into post-hoc and test-time methods. Large pre-trained model-based OOD detection methods focus on models such as vision language models or large language models, which are pre-trained on vast datasets and excel in a wide array of tasks. We discuss them in terms of whether they have access to a few examples, including zero-shot, few-shot and full-shot scenarios.

The remainder of this survey is organized as follows. We recap the related work of OOD detection in Sec. 2. Next, we summarize training-driven OOD detection approaches in Sec. 3, and introduce the training-agnostic OOD detection methods in Sec. 4. Then, in Sec. 5, we introduce large pre-trained model-based OOD detection. An overview of the evaluation metrics, experimental protocols, and applications is presented in Sec. 6. Following that, we discuss promising trends and open challenges in Sec. 7 to shed light on underexplored and potentially critical avenues. Finally, we conclude this survey in Sec. 8.

2 RELATED WORK

Anomaly Detection. Anomaly detection (AD) involves identifying data points, events, or observations that deviate significantly from the dataset's normal behavior [119]. These anomalies can indicate critical incidents, such as fraud [120], network intrusions [121], or system failures [122]. The process involves using statistical, machine learning, or deep learning methods to model normal behavior and detect deviations. Anomaly detection is crucial in domains like cybersecurity, finance, healthcare, and manufacturing, where recognizing unusual patterns quickly can prevent significant losses or improve operational efficiency. While both AD and OOD detection aim to identify unusual or

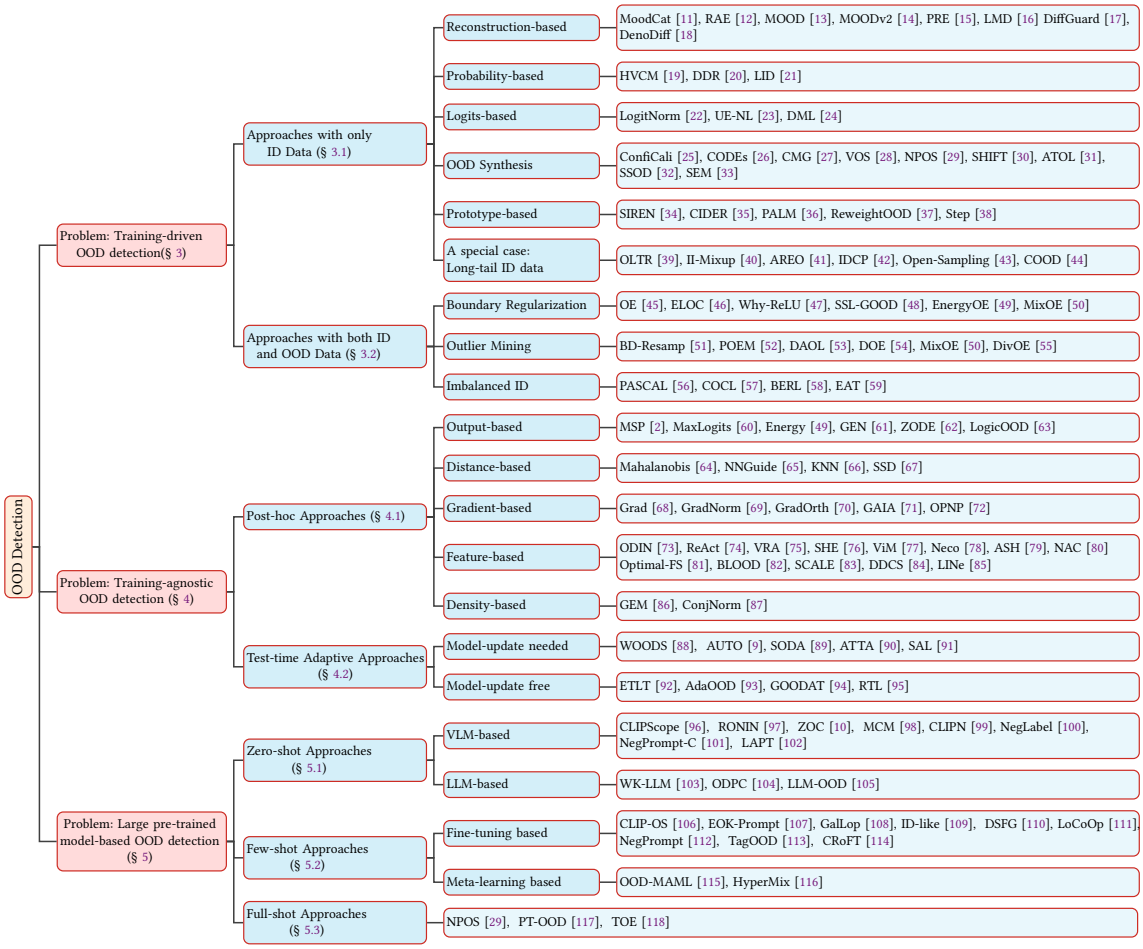

Fig. 1. Taxonomy of OOD detection problem scenarios and solutions.

unexpected data points, AD focuses on deviations within a single distribution (typically the training set), whereas OOD detection addresses differences between the training distribution and new, unseen inputs.

Novelty Detection. Novelty detection (ND) focuses on identifying new or unknown data points that a model has not seen during training [123]. This process is essential in situations where the system needs to adapt to evolving conditions or when it's crucial to flag new, unseen scenarios. Unlike AD, which seeks to find patterns that deviate from the norm, novelty detection aims to discover entirely new patterns. Applications include identifying new trends in social media [124], discovering new species in biological data [123], or detecting new topics in document streams [125]. Essentially, ND deals with surprises in familiar contexts, while OOD detection deals with data from unfamiliar contexts.

Open Set Recognition. Open set recognition (OSR) extends beyond traditional classification tasks by not only classifying known classes but also recognizing when input does not belong to any of the known categories [126]. This is crucial for real-world applications where the environment is dynamic, and the system encounters instances that were not present during training. Open set recognition is essential in fields like robotics [127], autonomous vehicles, and

image recognition systems [128], where encountering unknown objects or scenarios is common, and the system must be able to handle these gracefully. OSR is concerned with identifying new classes that are not present during training, which is generally achieved by dividing the categories of the same dataset into base classes and new classes, which means that new classes and base classes usually come from the same domain. In contrast, OOD detection focuses on identifying any data that is different from the training distribution, regardless of whether it belongs to a new class or a different domain altogether. Essentially, OSR is a subset of OOD detection, specifically aimed at classifying new class types within the same domain.

Outlier Detection. Outlier detection (OD) is the process of identifying data points that are significantly different from the majority of the data [129]. It is similar to AD but focuses more on the identification of individual data points rather than patterns. Outliers can arise due to variability in measurement, experimental errors, or genuinely novel variations. Applications include fraud detection, fault detection, and removing anomalous data from datasets to improve model accuracy [130]. Compared to OOD detection, OD is more of a transductive scenario, as it inherently has access to outliers. However, classical OOD detection only encounters outliers at the time of deployment.

Zero-shot learning. Zero-shot learning is a paradigm in machine learning where the goal is to recognize objects without having seen examples of these objects during the training phase [131–133]. The fundamental challenge in zero-shot learning is to effectively transfer knowledge from seen to unseen classes [131]. It has been primarily tackled by learning semantic relationships between classes through attributes or by embedding both seen and unseen classes into a shared semantic space. While both deal with unknowns during inference, zero-shot learning tries to classify new categories [134], whereas OOD detection flags data that is anomalous or unfamiliar to the training data distribution [64].

Selective Classification. Selective classification, also known as reject option classification, provides a mechanism for models to abstain from making a decision when they are not sufficiently confident in their predictions [135, 136]. Selective classification involves a model deciding when to make a prediction based on its confidence level, effectively choosing not to predict when uncertain [137]. On the other hand, OOD detection identifies data points that differ from the training distribution, aiming to flag them as unfamiliar to the model. While both methods manage uncertainty, selective classification deals with confidence in making predictions on ID data [136, 138], and OOD detection focuses on recognizing and handling data that is not represented in the training set.

3 PROBLEM: TRAINING-DRIVEN OOD DETECTION

In the *training-driven OOD detection* problem, researchers design the pre-training process to obtain models with superior OOD detection capabilities. Based on whether OOD data is accessible during training, we further divide methods under this scenario into two folds: training with only ID data and training with both ID and OOD data, as shown in Fig. 2.

3.1 OOD Detection Approaches with only ID Data

Overview. Given the ID data, approaches in this section train a model on them and aim to utilize the model for detecting OOD test samples while ensuring accurate classification of the ID data. Approaches with only ID data focus specifically on mining information from ID data, without explicitly relying on other information from real-world OOD data. We further differentiate these methods into the following four categories: *Reconstruction-based*, *Probability-based*, *Logits-based*, *OOD Synthesis*, and *Prototype-based*. Considering real-world requirements, we also delve into a specific scenario: *Long-tail ID data*.

Reconstruction-based. The reconstruction-based methodology offers a new research avenue by scrutinizing the discrepancy between sample representations before and after reconstruction, which relies heavily on the reconstruction

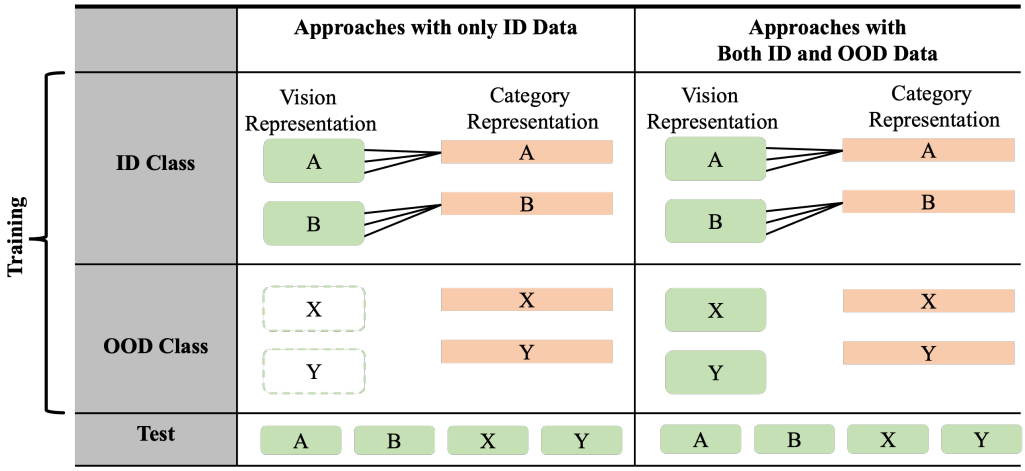

Fig. 2. Illustration of training-driven OOD detection approaches. Dashed borders indicate that they are not used in the specific phase. OOD images are excluded in the "Approaches with only ID data" on the left but are included in the "Approaches with both ID and OOD data" on the right. In both cases, OOD labels are not utilized.

performance of models. Fundamentally, the objective of the reconstruction task is to recuperate the inherent semantic content of the dataset, guided by a designated supervisory signal. This premise rests on the notion that OOD data intrinsically possess semantic characteristics that are incongruent with those of ID label information. By quantitatively assessing the extent of this semantic deviation, the model becomes equipped to discern OOD data with a high degree of precision. The approximate process of such methods is shown in Fig. 3.

MoodCat [11] masks random portions of input images and utilizes a generative model to synthesize new images based on the classification results, implementing strong constraints during the synthesis process. Similarly, Zhou [12] introduces an auxiliary module to extract activations of feature vectors, aiding the model in constraining the latent reconstruction space to filter potential OOD data. Following this, MOOD [13] and MOODv2 [14] leverage pretraining tasks based on masked image modeling, exhibiting significant advantages in learning the internal distribution of data. PRE [15] introduces the concept of normalized flow, combined with a penalty based on typicality to constrain reconstruction errors, which discerns differences between OOD and ID data well.

Recently, diffusion models have made remarkable strides in both training stability and the quality of generated images. Leveraging diffusion models to detect OOD data has become a new research direction. Graham et al. [18] address this issue by introducing DDPM, utilizing it for the reconstruction of noise-disturbed images. The model employs multidimensional reconstruction error for the identification of OOD data. The information bottleneck of this model can be externally regulated. Similarly, LMD [16] utilizes diffusion models for OOD validation. LMD disrupts data and then employs diffusion models to reconstruct images separated from the original manifold, distinguishing OOD data by comparing differences with the original manifold. DiffGuard [17] directly employs a pre-trained diffusion model for semantic mismatch guidance, aiming to leverage the diffusion model to amplify the disparity between reconstructed OOD images and the original images.

Probability-based. Research in probability-based direction aims to establish probability models to describe the distribution of training data. Through these probability models, suitable scoring functions are developed to compute the scores of test samples within the ID distribution, which reflect whether these samples belong to the ID distribution. In

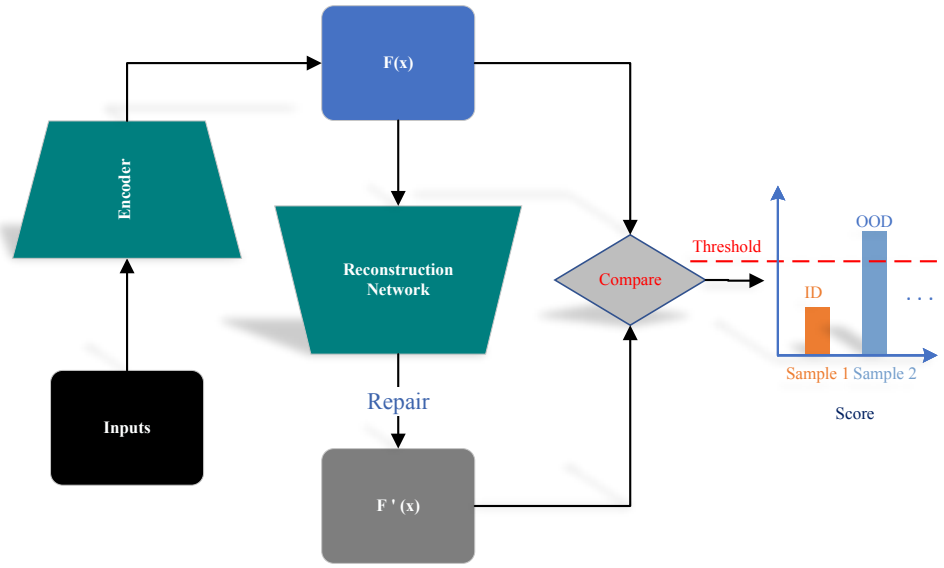

Fig. 3. Illustration of reconstruction-based OOD detection approaches. After extracting the feature representation $F(x)$ of the image through a neural network, it is then fed into a reconstruction network (such as VAE or DDPM) to obtain the reconstructed image representation $F'(x)$. By comparing the dissimilarity between the original and reconstructed images, we can identify OOD data, as OOD samples typically exhibit greater dissimilarity before and after reconstruction.

this domain, Li et al. [19] model each ID category during training using multiple Gaussian mixture models, and during the prediction phase, it combines Mahalanobis distance metrics to assess the likelihood of anomalous classes. Huang et al. [20] address this issue by introducing two regularization constraints. The density consistency regularization aligns the analytical density with low-dimensional class labels, and the contrastive distribution regularization helps separate the density between ID and OOD samples. Furthermore, LID [21] introduces a new detection criterion for the OOD detection paradox in the context of data generation by deep generative models. It measures whether data should be classified as in-distribution by estimating the local intrinsic dimension (LID) of the learned manifold of the generative model, when the data is assigned a high probability and the probability mass is non-negligible.

Logits-based. These algorithms primarily focus on the predictions of neural networks, particularly on Logits, which are the results of the neural network's output layer. Logits typically represent the model's confidence or probability for each category. LogitNorm [22] proposes a method using logit normalization to enforce a constant vector norm on logits during training to mitigate issues of model overconfidence. Similarly, UE-NL [23], derived from Bayesian networks, normalizes logits while simultaneously learning embeddings and uncertainty scores. It adjusts the learning intensity between samples during training, rendering the learning process robust. DML [24] challenges the potential hindrance to OOD detection performance. Drawing on empirical insights, DML enhances OOD detection performance by decoupling logits and balancing individual components, thereby mitigating the impact of each attribute on the results. It enhances OOD detection performance by decoupling logits and scenario MaxNorm as a parameter, thereby balancing the influence of each attribute on the outcome.

OOD Synthesis. In the task of OOD detection, it is theoretically believed that incorporating the features of OOD data during model training can enhance the OOD detection performance. Due to the challenges in acquiring distribution information for OOD samples, some methods employ ID data to estimate the distribution of OOD data. This is done to simulate the scenario in which a model encounters OOD data in real-world situations.

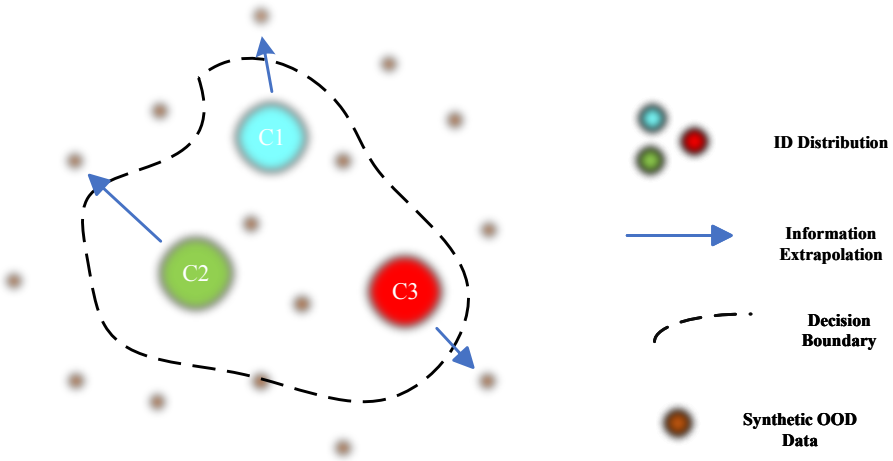

Fig. 4. Illustration of OOD synthesis approaches. In the absence of real-world OOD data, utilizing extrapolation from ID data to identify suitable OOD data and then training the model using the method of exposing anomalies enhances the capability of OOD detection. The orange dots in the figure represent virtual OOD data synthesized using ID data.

Lee et al. [25] add additional regularization terms by incorporating two supplementary terms into the original loss function. The first term compels the classifier to increase uncertainty for OOD samples, while the second term encourages the GAN to generate appropriate OOD samples. Specifically, a joint training scheme is designed to select suitable boundary samples in the low-density regions of ID samples. Similarly, VOS [28] employs adaptive synthesis of anomaly values. It samples anomaly values from the low-likelihood regions of class-conditional distributions in the feature space, eliminating the need for reliance on external data. This enables the model to synthesize OOD data for predictions. It is worth noting that VOS supports adaptive outlier synthesis, making it easily applicable to any ID data without the need for manual intervention. Noticing the additional distributional assumptions posed by previous methods for OOD data, Some researches approach the problem from another perspective. For example, NPOS [29] contends that, rather than modeling the feature space of outliers as a parameterized Gaussian distribution, not making any distribution assumptions about ID embedded data would endow the model with robust flexibility and generality. it selects boundary points through non-parametric density estimation based on nearest neighbors, with these boundary points situated between ID and OOD data. Additionally, SSOD [32] introduces self-supervised sampling to implicitly generate OOD data. It directly samples natural OOD signals from the backgrounds of ID data images, thus overcoming limitations introduced by biases during the OOD synthesis stage.

In contrast to the implicit generation methods of OOD samples mentioned above, some approaches take a different route by directly leveraging explicit construction of OOD samples from ID samples for learning. To acquire OOD samples, CODEs [26] initially generate seed OOD examples by slicing and stitching samples from internal distributions of different categories. These examples are then inputted into Chamfer GAN for distribution transformation, yielding

high-quality OOD samples. CMG [27] then generates pseudo OOD data by providing class embeddings mixed as conditions to a Conditional Variational Autoencoder (CVAE), and subsequently utilizes this data to fine-tune the classifier for OOD detection. SEM [33] introduces a problem setting for full-spectrum OOD detection, where negative samples are generated using Mixup operations during training. It then leverages both high-level semantic information and low-level non-semantic features, such as style, to identify the source of the data. SHIFT [30] proposes a direct synthesis of OOD image samples based on training examples. It achieves this by employing CLIP to eliminate ID object regions within the training samples. The latent diffusion model is then utilized to substitute these regions with authentic features, all the while taking into account the contextual background. This methodology thereby establishes the model's capability for rejection. However, ensuring the quality of generated data is often challenging. Furthermore, the quality of the generated data itself may have inherent flaws. ATOL [31] introduces an auxiliary task, wherein both auxiliary ID and auxiliary OOD data coexist. In a low-dimensional latent space, distinct regions are manually sought for both auxiliary ID and auxiliary OOD data, ensuring non-overlapping characteristics in the input space. Subsequently, data generation by the generator guarantees the non-overlapping properties of the input space. Finally, reliability is upheld by aligning genuine ID data with auxiliary ID data, effectively alleviating issues related to erroneously generated OOD instances.

Prototype-based. During the model training process, prototype-based OOD detection methods aim to model the ID data using prototypes to learn the common distribution characteristics of the ID data. In the testing phase, the model measures the differences between the sample and class-level prototypes to determine the category of the sample. The general procedure of these methods is roughly illustrated in Fig. 5.

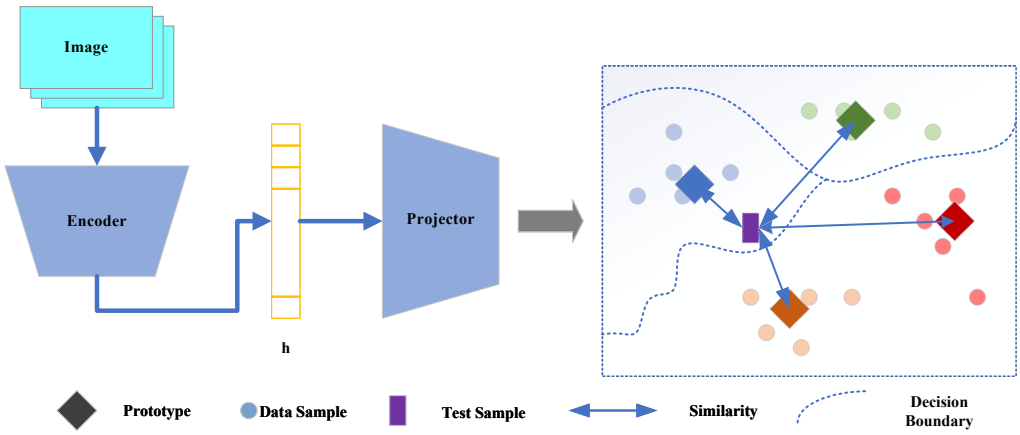

Fig. 5. Illustration of prototype-based OOD detection approaches. The algorithm for OOD detection based on prototype learning roughly consists of three network structures: 1) **Encoder**. This structure is responsible for propagating sample data through a neural network to obtain feature representations h. 2) **Projector**. Structures of this type project the image features encoded in h to a new representation space. 3) **Prototype Learning**. In the new representation space, prototypes representing each class are learned, with each class learning its corresponding prototype center. During testing, the likelihood of anomalous data is measured by comparing the similarity between test samples and prototypes of each ID class.

SIREN [34] first models the distribution of ID data using the von Mises-Fisher(vMF) model, which enables representing each class as a compact cluster, i.e., a class prototype. Generally, the vMF distribution modeling formula can be represented as: $p_D(z; p_k, k) = Z_D(k)exp(kp_k^T z)$, where p_k is the k-th prototype with unit norm, $\kappa \geq 0$ represents the

concentration around the mean, and $Z_D(\kappa)$ is the normalization factor. In the prototype-based approach, an embedding vector z is assigned to class c with the following normalized probability:

$$p(y = c|z; (P_j, k_j)_{j=1}^{C}) = \frac{Z_D(k_c)exp(k_c\mu_c^T z)}{\sum_{j=1}^{C} Z_d(k_j)exp(k_j\mu_j^T z)}, \tag{1}$$

where $c \in \{1, 2, ...C\}$. Additionally, within the loss function, it enforces alignment between the embedding vectors of ID samples and class prototypes, constraining each ID class sample. This parametric OOD score can be directly obtained after training, without requiring separate estimation. CIDER [35] builds upon SIREN by jointly optimizing two losses to enhance data discriminability, encouraging the maximization of angular distances between prototypes of different classes and the internal compactness of prototypes within the same class. The optimization process during model training pertains to the prototypes of classes. ReweightOOD [37] argues that optimizing non-class data impedes achieving clear class separability, while focusing on fewer class data makes it challenging to achieve lower MSE scores. To address this, they propose a re-weighting optimization strategy to balance the significance of different losses. Although the ideas behind Step [38] are different, in the context of semi-supervised tasks, it essentially generates clusters of unlabeled ID and OOD samples through a contrastive learning process, which is conceptually similar to prototype learning.

However, PALM [36] observes that modeling each class with a single prototype may fail to capture the diversity of internal information within the data. PALM introduces a mixed-prototype strategy to optimize the modeling process, utilizing a strategy of multiple prototypes to learn informative representations for each class. By simultaneously learning class-level prototypes and contrasting inter-class prototypes, it optimizes prototype-level intra-class compactness and inter-class discriminability within the loss function.

A special case: Long-tail ID data. ID data in real scenarios may present a long-tailed distribution due to the difficulty of collection and frequency of occurrence. This imbalance will strongly affect the performance of OOD detection. Many approaches are proposed to address the challenge of ID imbalance and enhance OOD detection capabilities.

OLTR [39] addresses the robustness of tail recognition by associating visual concepts between the head and tail embeddings. It accomplishes this by dynamically adjusting embedding norms through the utilization of visual memory. The approach involves two components to enhance the robustness of long-tailed data: one utilizes direct features from standard embeddings, while the other employs memory features that store discriminative centroids derived from direct features, thereby augmenting the direct features of the images. Mehta et al. [40] propose a tailored complex subset mixing strategy designed specifically for handling medium-to-tail classes. This method creates mixed samples by pairing two independent samples from different tail categories and calculates the distance between the mixed sample and prototypes of specific classes. The mixing strategy is ultimately combined with prototype learning to effectively address the challenges posed by long-tail scenarios. AREO [41] introduces a method to quantify sample uncertainty through evidential learning, akin to confidence. The entire training process is governed by an innovative multi-scheduler learning mechanism, dynamically adjusting training parameters based on the importance of different classes to ensure the model focuses on the features of minority classes, thus balancing the disparities between majority and minority classes. Overall, AREO identifies higher uncertainty in certain samples during training and dynamically adjusts parameters to mitigate overfitting to these samples.

Existing OOD detection methods predominantly assume a uniform distribution of probabilities from OOD to ID. In the presence of class imbalances, Jiang et al. [42] propose an alternative strategy to enhance current OOD detection methodologies. This involves recalibrating the OOD score based on the class prior distribution of the ID and the KL

divergence from the output of a pre-trained model. This approach signifies the potential to augment the robustness of OOD detection beyond the achievements of previously optimized methodologies, further fortifying the overall performance. Additionally, Open-Sampling [43] leverages noisy labels from the OOD dataset to rebalance the class priors of the ID training dataset. These labels are sampled from a predefined distribution that complements the original class prior distribution. COOD [44] uses the supervised model to combine individual OOD measures into a single ensemble, similar to the concept of random forests. This approach addresses the limitations of individual OOD methods and can also overcome issues related to data imbalance.

3.2 OOD Detection Approaches with Both ID and OOD Data

Overview. In some known deployment scenarios, real OOD data can be easily collected at a low cost. Some methods based on this assumption focus on how to use OOD data for better detection performance. Differing from methods involving OOD Synthesis, in these research directions, models have access to real-world OOD data during the training phase. The primary focus in such problems is on optimizing the model's decision boundary, rather than the OOD data itself. Due to the introduction of real OOD information, the boundary of these two categories will be accurately calculated.

Boundary Regularization. The Boundary Regularization class of methods belongs to the traditional Outlier Exposure (OE) approaches. The central idea of Hendrycks et al. [45] and Hein et al. [47] are to fully leverage OOD data to optimize the model's decision boundary, thus achieving OOD detection. Proponents of this concept can utilize auxiliary anomaly datasets to enhance the OOD detector, enabling it to generalize and detect anomalous information not encountered during training. The central idea of this method can be grasped from Fig. 6.

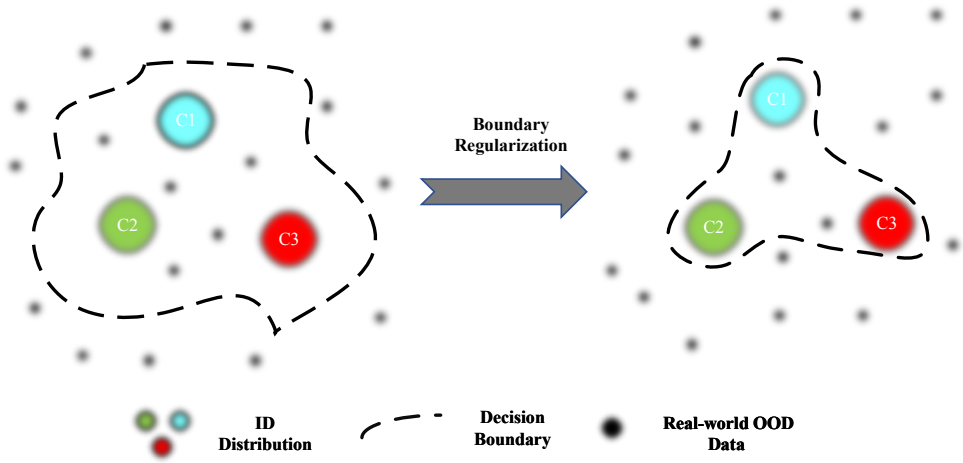

Fig. 6. Illustration of boundary regularization approaches. The key to model optimization strategy lies in the decision boundary of the model classification. Boundary regularization utilizes existing known ID data (the colored dots) and some OOD data (black dots) to optimize the potential decision boundary, compacting its classification boundary with abnormal data as tightly as possible without affecting the classification of ID data.

Specifically, given a model f and the original loss function, the model training process aims to minimize the objective:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{in}}}\left[\mathcal{L}(f(x),y)+\lambda\mathbb{E}_{x'\sim\mathcal{D}_{\text{out}}^{\text{OE}}}\left[\mathcal{L}_{\text{OE}}(U,f(x))\right]\right], \tag{2}$$

over the parameters of f, where x' represents auxiliary anomaly data and U denotes the uniform distribution. $\mathcal{L}_{\text{OE}}$ represents the cross-entropy loss with respect to U. The fundamental purpose is to compel the model to optimize the OOD data distribution to a uniform distribution, a principle that is universal in OE-type approaches. The specific design of $\mathcal{L}_{\text{OE}}$ can depend on other task requirements and the chosen OOD score. This design can utilize a maximum softmax probability baseline [2] detector to detect anomalous data. Compared to traditional softmax scores, EnergyOE [49] builds upon OE by leveraging energy scores for better discrimination between ID and OOD samples, and it is less prone to issues of overconfidence. Specifically, its calculation formula:

$$E(x;f) = -T \cdot log \sum_{i}^{K} e^{f_i(x)/T}, \tag{3}$$

where the temperature coefficient T is used and $f(x)$ denotes the discriminative neural classifier $f(x) : \mathbb{R}^D \to \mathbb{R}^K$, which maps an input $x \in \mathbb{R}^D$ to K real-valued logits.

Mohseni et al. [48] train a model using a self-supervised approach, optimizing the objective function for unlabeled OOD samples using pseudo-labeling to generalize OOD detection capabilities. Vyas et al. [46] similarly employ self-supervised training of the classifier. Unlike the OE approach, its aim is to find a gap between the average entropy of OOD and ID samples. MixOE [50] takes into account the beneficial effect of subtle OOD samples on enhancing the generalization ability of OOD detection. Its main idea is to mix ID and OOD data samples to broaden the generalization of OOD data. Training the model with these outliers can linearly decrease the prediction confidence with inputs from ID to OOD samples, explicitly optimizing the generalization ability of the decision maker.

Outlier Mining. The traditional OE concept assumes the existence of ID input $\mathcal{D}_{\text{in}}$ and OOD input $\mathcal{D}_{\text{out}}$, both independently and heterogeneously distributed, originating from different sources. However, this premise cannot be fully guaranteed in the current training process due to potential noise in the training OOD data. Outlier Mining differs slightly from the traditional OE approach in that, although it also utilizes real-world OOD samples to address the issue, it focuses on identifying the optimal selection within the existing OOD data. The main process is depicted in Fig. 7.

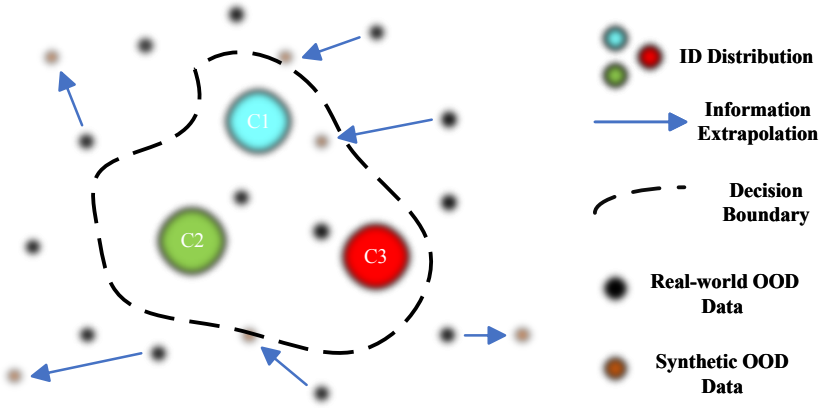

Fig. 7. Illustration of outlier mining approaches. Due to the limitation of whether real OOD data exists, methods like anomaly mining allow the model to access real OOD data during training. The black dots in the figure represent real OOD data sampled from other datasets, which do not overlap with the ID categories. Utilizing such OOD data to extrapolate more representative OOD data (orange dots) is a solution to this type of problem.

POEM [52] raises methodology leans towards excavating anomalies that carry more representative characteristics. Li and Vasconcelos [51] propose a method of data resampling to acquire representative outlier data for training, assigning higher priority score reweighting to hard negative instances. Iterative optimization, leveraging adversarial principles, selects target data. POEM employs posterior sampling to unearth anomalies with elevated boundary scores from an extensive auxiliary dataset, facilitating a nuanced comprehension of the intricacies within OOD samples. DAOL [53] takes the disparity between authentic data and auxiliary OOD data as a starting point to formulate an OOD dataset. Utilizing the Wasserstein ball, it models the distribution of all OOD data, selecting the most challenging OOD data within the ball for training.

Beyond solely relying on raw data, another direction in addressing this issue involves synthesizing representative outlier data by utilizing authentic OOD data through information extrapolation. DivOE [55] introduces a novel learning objective to alleviate challenges associated with limited auxiliary OOD datasets. It achieves this by adaptively inferring and learning information from surrogate OOD data through the maximization of differences between generated OOD data and original data, given the specified anomalies. This adaptive inference extends to a broader spectrum, addressing the limitations imposed by a finite auxiliary OOD dataset. Moreover, DOE [54] introduces a Min-Max learning strategy to identify the most challenging OOD data for a synthetic model. Through model perturbation, the data is implicitly transformed, and the model continues learning from this perturbed data to improve its robustness.

Imbalance. Given practical requirements, research is progressively increasing regarding scenarios where ID data is imbalanced during training yet still provides OOD data information. The concept behind PASCAL [56] is to segregate tail-end data and OOD data solely through contrastive loss during training phase, aiding the model in better distinguishing between the two. To address the issue of model confusion between OOD samples and tail-end class data, COCL [57] incorporates a learnable tail class prototype during the training process. This prototype serves to bring tail-end samples closer while distancing them from OOD data, thereby mitigating the model's bias towards OOD samples. Choi et al. [58] take a different approach, suggesting that factors affecting the performance of OOD detection may be related to the imbalance in the cross-class distribution of auxiliary OOD data. Consequently, it proposes an energy regularization loss, specifically regularizing auxiliary samples from the majority classes to address the issue of class imbalance in OOD data. EAT [59] utilizes dynamically assigned virtual labels during the model training process to train OOD data, thereby expanding the classification space.

Despite the success and considerable attention received by methods like Outlier Exposure in the research community, there are voices questioning the essence of allowing access to OOD data during training. Nevertheless, concerns are raised that the superior classification performance observed in certain datasets may not necessarily translate to competitiveness in real-world deployment, challenging the original intention of OOD detection.

4 PROBLEM: TRAINING-AGNOSTIC OOD DETECTION

Training-agnostic OOD detection places a primary emphasis on adaptation strategies during test time, as opposed to the focus on classifier performance seen in training-driven OOD detection. Based on whether they rely on dependencies among test data, methods are categorized into two types: post-hoc and test-time adaptive approaches, as illustrated in Fig. 8. Post-hoc methods compute results for individual samples independently, unaffected by changes in other samples. In contrast, test-time adaptive methods exhibit dependencies among test samples.

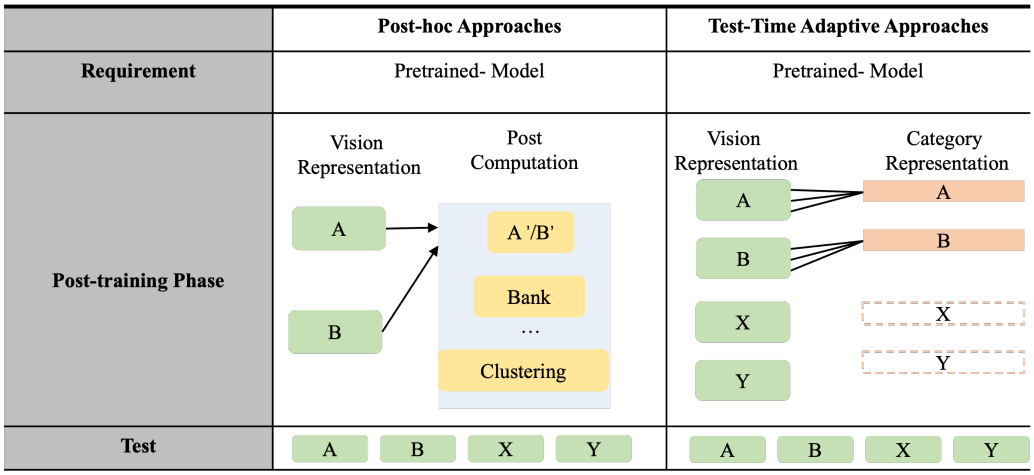

Fig. 8. Illustration of training-agnostic OOD detection approaches. Both methods require access to a pre-trained model. Post-hoc approaches do not involve any operations during the post-training phase, while test-time adaptive approaches necessitate adaptation based on the samples encountered during testing. "A'/B' " means that the original features are deformed; Bank means that some samples are stored; Clustering means that clustering is performed for ID images.

4.1 Post-hoc OOD Detection Approaches

Overview. Given a well-trained model, this problem scenario involves utilizing only the intermediate results computed by the trained model during testing, without modifying any parameters of the model, to accomplish the OOD detection task. Post-hoc methods are favored for their lightweight nature, low computational costs, and the fact that they require minimal modifications to the model and objectives. Its main objective is to construct an effective scoring function that can accurately reflect the behavior of ID data. These characteristics make them highly desirable for convenient and straightforward deployment in practical scenarios.

Post-hoc approaches are categorized into five types: *Output-based*, *Distance-based*, *Gradient-based*, *Feature-based* and *Density-based*. Recent work on this type of problem has some recent progress. A summary of the key factors involved in such methods is given in Table 1.

Output-based. Algorithms based on output primarily aim to explore the latent representations of the output from the intermediate layers of neural networks, which include logits and class distributions, among others. MSP [2] was the first to employ the maximum softmax value to validate OOD detection effectiveness. For OOD samples, their output probability distribution tends to be closer to a uniform distribution, demonstrating the model's inability to correctly classify the category. In contrast to the MSP method, MaxLogits [60] detects anomalies by comparing the maximum logit value in the logits vector output by the neural network. Logits, representing the model's confidence in each category, are the neural network's output before the softmax layer, without undergoing softmax transformation. Meanwhile, Energy [49] introduces the Helmholtz Free Energy, which theoretically aligns with the input probability density and is less susceptible to overconfidence issues. GEN [61] introduces the concept of generalized entropy and directly utilizes Bregman divergence to compute the statistical distance between the model's probability output and uniform distribution, aiming to identify OOD data. Additionally, leveraging sufficient prior knowledge might be a viable solution. ZODE [62] perform predictions on samples across multiple pre-trained models simultaneously to determine

Table 1. Comparison of key components in OOD detection methods.

Method	Type	Space			
		feature	logit	gradient	probability
MSP [2]	Output-Based				✓
Maxlogits [60]			✓		
Energy [49]			✓		
GEN [61]		✓			✓
Mahalanobis [64]	Distance-Based	✓			
NNGuide [65]		✓			
KNN [66]		✓			
SSD [67]		✓			
Grad [68]	Gradient-Based			✓	
GradNorm [69]		✓		✓	✓
GradOrth [70]		✓		✓	
GAIA [71]		✓		✓	
OPNP [72]		✓		✓	
ODIN [73]	Feature-Based	✓		✓	✓
ReAct [74]		✓			
VRA [75]		✓			
SHE [76]		✓			
Vim [77]		✓	✓		
Neco [78]		✓	✓		
ASH [79]		✓			
NAC [80]		✓			
Optimal-FS [81]		✓	✓		
BLOOD [82]		✓			
SCALE [83]		✓			
ConjNorm [87]	Density-based	✓	✓		
GEM [86]		✓	✓		✓

whether multiple models can identify OOD samples, using this as a basis to distinguish between data. LogicOOD [63] present a novel approach that uses first-order logic for knowledge representation to perform OOD detection. This reasoning system uses prior knowledge to infer whether an input is consistent with the prior knowledge about the training distribution. It is particularly user-friendly in terms of interpretability, as it allows for comparing the output of samples against a knowledge base to determine whether they belong to the ID data.

Distance-based. Another approach in OOD detection research focuses on measuring statistical distance metrics. Mahalanobis [64] is typically computed by calculating the distance between the feature vector and its mean. Specifically, for each class, we compute the mean and covariance matrix of its feature vectors. During testing, it calculates the Mahalanobis distance between the feature vector and the mean of each class. SSD [67] essentially utilizes the Mahalanobis distance. After being trained on unlabeled ID data via self-supervised representation learning, it employs the Mahalanobis distance as a statistical measure for classification using the pre-trained model. In comparison, while the Mahalanobis distance makes strong distributional assumptions about the data, KNN [66] explores the effectiveness of non-parametric nearest-neighbor distance for OOD detection. By measuring the k-nearest neighbor distance between input embeddings

and training set embeddings, a threshold is designed to determine whether the data belongs to the ID data. NNGuide [65] takes a step further in the direction of granularity by combining the idea of KNN. It assigns weights before the traditional OOD Score, depending on the nearest neighbor distance between the sample and the embeddings in the training set.

Gradient-based. Grad [68] research indicates that gradient-based methods also contribute to OOD detection by quantifying model uncertainty through the gradients propagated in reverse during backpropagation. If the input sample is an ID sample, the gradients of the model with respect to these samples tend to be relatively small and stable. Conversely, the gradients are typically larger or more irregular for OOD samples. GradNorm [69] posits that the gradient magnitude for ID data surpasses that of OOD data. Leveraging this observation, it employs gradient vector norms, computing KL divergence through softmax output and uniform distribution backpropagation to detect OOD data. In contrast, GradOrth [70] adopts an alternative perspective, recognizing that crucial features of OOD data reside in a low-rank subspace. Consequently, it shifts focus to calculating gradient projection norms in this subspace to identify OOD data. GAIA [71] employs a combination of gradient anomaly checks and aggregation. It enables the model to interpret indeterminacy attributively, introducing channel-wise average abnormality and zero-deflation abnormality without prior knowledge to gauge the extent of data distribution variations. The OPNP [72] method discovers that the OOD detection capability of the model is highly sensitive to parameters and activated neurons that are close to zero. Therefore, this method utilizes a pruning behavior for parameters and neurons to remove those leading to overfitting, thereby enhancing the model's generalization ability.

Feature-based. Another avenue of research into OOD detection involves the impact of neural network intermediate variables relative to the final prediction. Inspired by adversarial examples, ODIN [73] experimentally discovered that adding small perturbations to the input image features can more effectively detect OOD data. The perturbation formula is as follows:

$$\tilde{x} = x + \varepsilon sign(\nabla_x log max_c p_c(x)), \tag{4}$$

where the parameter ε is the perturbation magnitude. For a given input x, compute its logit output $p_c(x)$. Therefore, ReAct [74] focuses on the high activation values in the intermediate results of the model. These activation values do not affect the model's classification, but truncating high activation values can significantly improve OOD detection performance. VRA [75], an extended iteration of ReAct [74], builds upon the premise of ReAct which truncates only high activation values. However, VRA posits that this might not be the optimal solution and, therefore, employs a variational approach to seek an optimal solution. It utilizes piecewise functions to emulate suppression or amplification operations, aiding the model in recognizing anomalous data. SHE [76] converts the output of the penultimate layer into a storage pattern, utilizing Hopfield energy [139] for OOD sample detection through a pattern of storing before comparison. It employs a storage mechanism to calculate the average logits for each class, preparing this simple storage pattern for subsequent OOD detection tasks. Drawing inspiration from experience, DDCS [84] adapts to select suitable channels for data classification after correcting each channel in the neural network. These channels are evaluated based on inter-class similarity and variance to measure their discriminative power for ID data. LINe [85] similarly emphasizes neuron outputs at the feature level. It uses Shapley value pruning to select only high-contribution neurons for prediction while masking out the remaining input data, thereby reducing the influence of irrelevant outputs.

Feature shaping is emerging as a prominent area of research focus. This methodology entails the refinement of intermediate representations—specifically, the intermediate features—within the forward propagation phase of the model. The salient advantage of this approach lies in its non-disruptive nature concerning the original classification results, coupled with its simplicity and efficacy. Building upon this foundation, ViM [77] simultaneously considers the

roles of features, logits, and probabilities, constructing virtual logits to aggregate information for decision-making. By decomposing the features of the penultimate layer in a neural network, it identifies the null space that is irrelevant to classification but exhibits exceptional performance in OOD detection. Its computation formula can be expressed in the following form:

$$-\alpha\|z^{P^{\perp}}\|^2 + \text{LogSumExp} f(z), \tag{5}$$

where α is a scaling constant, computed by the model. Here $z = z^P + z^{P^{\perp}}$ and $z^{P^{\perp}}$ is the projection of z to $P^{\perp}$. And it have $Wz^{P^{\perp}} = 0$. Additionally, *LogSumExp* represents the computation process of the energy function [49], and $f(z)$ represents the logit output of the model. The first term here represents virtual logits, while the second term represents the score of the energy function. Neco [78] subsequently reveals the prevalent phenomenon of neural collapse in contemporary neural networks, impacting OOD detection performance. The observation of orthogonal trends between ID data and OOD data features is leveraged to differentiate OOD data. ASH [79] is a straightforward method for dynamic activation modification where a substantial portion of activations in samples during later stages of model training is either removed or slightly adjusted. NAC [80] introduces a measure of neuron activation coverage, based on the premise that if a neuron in a neural network is rarely activated, this state may indicate a higher likelihood of OOD data. By quantifying this statistical property, NAC aims to distinguish between ID and OOD data. Based on the existing literature on *feature shaping*, Zhao et al. [81] utilize a piecewise constant shaping function to partition the feature domain into disjoint intervals, estimating a scalar within each interval as an approximation. As the interval width approaches zero, an approximation of the maximum logits can be obtained. The goal of BLOOD [82] is to smooth the representations of intermediate layers in neural networks for predicting OOD data. This study found that, compared to OOD data, ID data exhibits smoother variations in the intermediate layer representations of neural networks. Leveraging this characteristic, new statistical measures can be designed to discriminate against anomalous data. In addition, another line of work focused on neuron activation pruning, which is a new activation shaping scheme proposed based on the research foundation of ASH. SCALE [83] emphasizes scaling as a crucial metric for evaluating samples, and it similarly finds significantly lower pruning rates for OOD data. Therefore, the proposed reshaping of intermediate layer tensors can effectively enhance detection performance.

Density-based. Recent advancements in density-based OOD detection models have demonstrated substantial performance gains, predicated on the models' ability to accurately capture and understand the intrinsic characteristics of the true data distribution. For example, GEM [86] models the feature space of ID data as a class-conditional multivariate Gaussian distribution. Under this assumption, it designs new statistical metrics to validate the model's performance. Using a modeling approach based on Gaussian mixture models, GEM Score is aligned with the true log-likelihood for capturing OOD uncertainty. However, while GEM strictly relies on Gaussian assumptions, recent works ConjNorm [87] introduce a novel framework based on Bregman divergence, extending the consideration of data distributions to encompass an exponential family of distributions. By redefining density functions for data, this model has a broader range of applications.

4.2 Test-Time Adaptive OOD Detection Approaches

Overview. Test-time adaptive methods, which employ a classifier trained on the training set, strive to utilize test data, either the complete test set or a series of unlabeled mini-batches, to enhance OOD detection performance through model adaptation. Test-time adaptive approaches are based on the theoretical insight [140] that detecting OOD samples using only ID samples without any additional knowledge is impossible. These methods can be divided into two categories

according to whether the model be modified during the testing time: *Model-optimization-based* and *Model-optimization-free*. Both of them undergo a post-training phase, during which the trained model can be adapted, regardless of whether it is updated.

Model-optimization-based. Model optimization-based methods enhance the trained model by leveraging unlabeled data during the post-training phase.

A line of these methods [88, 91] advocate for harnessing a combination of unlabeled ID and OOD data, termed "wild data", which is plentiful and readily available in the real-world scenario, to OOD detection. The motivation of WOODS [88] is to clean the wild data to get reliable OOD candidates, and then the model regularization can be performed with the knowledge of them. Afterwards, to understand the role of wild data in OOD detection, SAL [91] explains how they help OOD detection from the lens of separability and learnability. Notably, wild data is not entirely equivalent to test OOD data if wild OOD data is not from the test datasets, inevitably leading to incorporating unintended information into the model. In addition, the extra training requirement in WOODS, thought necessary, incurs significant costs and is generally undesirable.

Another line of model-optimization-based approaches draws inspiration from the semi-supervised-learning (SSL) techniques [141], aiming for a more efficient, lightweight training process during the post-training phase. Pseudo-labeling [142] is a simple but effective way to label the test data, which enhances learning from test unlabeled data. The method proposed by Yang et al. [9], termed AUTO, employs only pseudo-OOD data to refine the model. The role of pseudo-ID data in AUTO is to mitigate catastrophic forgetting with a semantically-consistent objective, thereby maintaining the accuracy of ID classification. In contrast, ATTA [90] and SODA [89] harness both pseudo-ID and pseudo-OOD data to refine the trained model. SODA employs a dual-loss approach to tackle pseudo-ID and pseudo-OOD data simultaneously, while ATTA distinguishes them with different weighting techniques.

Model-optimization-free. Modifying the original trained model is infeasible in certain security-sensitive scenarios. Therefore, methods enabling test-time adaptation without requiring model updates, termed "model-optimization-free" (MOF) techniques, are increasingly garnering interest. These approaches enhance the utilization of test data by either memorizing it or incorporating additional modules on top of the original model.

Both ETLT [92] and GOODAT [94] retain the integrity of the original model by training an add-on module to adjust the OOD score, rather than altering the original model itself. ETLT observes a linear correlation between the feature representation and the OOD score of a test input. In other words, for a given image, the pair consisting of its feature and OOD score *(feature, OOD score)* exhibits a linear correlation. Furthermore, the aforementioned pairs of ID and OOD data is linearly separable. Based on the observation, ETLT proposes to learn a linear regression module trained from the *(feature, OOD score)* pair. The availability of complete test datasets may not always be feasible, therefore, Fan et al. [92] also provide an online variant to ensure safer deployment. Similarly, GOODAT develops an add-on named graph masker, designed specifically for graph data. It integrates GIB-boosted losses and employs it as the metric for OOD scoring. In contrast, AdaOOD [93] avoids any additional training burden through a non-parametric k-nearest neighbours approach. The core principle behind AdaOOD is the maintenance of a memory bank, similar to the approach taken by AUTO.

Online v.s. offline. Most post-hoc methods [92, 143] have traditionally emphasized the offline scenario, where OOD detectors remain static and fixed after deployment. In contrast, the majority of test-time methods [90, 93] adopt the online scenario to obtain the decision boundary dynamically, minimizing the risk of incorrect OOD predictions at each time step.

More challenging scenario. In the context of test-time OOD detection scenarios, some scholars have proposed more challenging configurations that demand a higher level of capability from the models. MOL [144] introduces a more

realistic problem scenario, namely Continuous Adaptive Out-of-Distribution (CAOOD) detection, aimed at addressing the challenge of constantly changing ID and OOD distributions in the real world. The meta-learning approach is employed to swiftly adapt models in response to the complexities encountered in various scenarios in MOL.

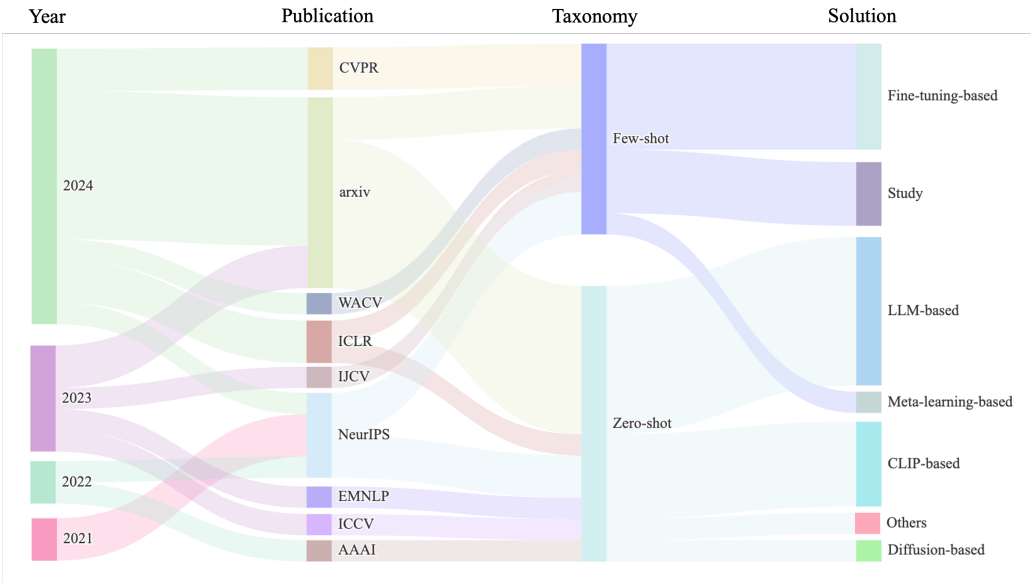

Fig. 9. Overview and trends in LPM-based OOD Detection Methods.

5 PROBLEM: LARGE PRE-TRAINED MODEL-BASED OOD DETECTION

Large pre-trained models have showcased remarkable performance in numerous downstream ID classification tasks, but their potential in OOD detection tasks remains a less-explored area. Recent research [145] highlights a correlation between higher ID classification accuracy and better OOD detection performance. Consequently, large pre-trained model-based OOD detection problem comes naturally. In recent years, large pre-trained models of various types, including single-modal (ViT [146], BERT [147], Diffusion [148]), visual language models (VLMs) (CLIP [149], multi-modal Diffusion [150] ALIGN [151],), and large language models (LLMs) (GPT3 [152]), have been increasingly utilized for OOD detection tasks, as shown in Fig. 9. Leveraging the powerful representational capabilities of large pre-trained models has further relaxed the constraints of OOD detection tasks, leading to a focus on more challenging and realistic scenarios, which has emerged as a new hotspot. Given the number of ID shots exposed to the large pre-trained model, large pre-trained model-based OOD detection can be classified into Zero-shot, Few-shot, and Full-shot OOD detection, as shown in Fig. 10. The performance evaluations of several relevant competitive methods are summarized in Table 2 to provide an understanding of the performance level of OOD detection in this area.

5.1 Zero-shot OOD Detection Approaches

Overview. Given the large pre-trained model and ID class names, we undertake the same task as the OOD detection, precisely detecting OOD data to abstain from prediction and accurately classifying ID data. Note that we do not need to

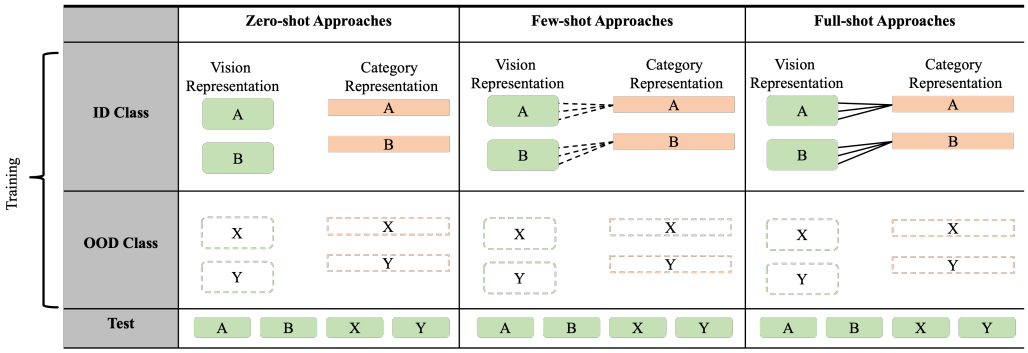

Fig. 10. Illustration of large pre-trained model-based OOD detection approaches. In the training phase, zero-shot approaches require only category labels of ID classes. Few-shot approaches need a subset of images of each ID class along with the category labels (indicated by dashed lines). Full-shot approaches utilize both the category labels and all images of each ID class. None of these approaches use labels or images of OOD categories.

have access to ID images and rely solely on textual category knowledge. It's important to clarify that "ID" here refers to the ID of the specific downstream task, not the dataset ID data during pre-training.

VLMs-based. Existing research on zero-shot OOD detection using VLMs can be categorized into two main approaches based on which VLMs are utilized as a basic backbone. **Diffusion-based.** Traditional generative approaches, such as diffusion models [148], capture the ID distribution effectively. Subsequently, these methods are used to identify OOD samples by assessing the likelihood that a given test input is derived from an OOD source. In a recent study in RONIN [97], the diffusion model has been applied to achieve OOD object detection, which utilizes the diffusion model to generate ID inpainting, while the CLIP model is employed to calculate the similarity score. **CLIP-based.** Given the remarkable proficiency of CLIP in correlating texts and images, a substantial number of researchers have ventured into zero-shot OOD detection leveraging CLIP. Typically, in the zero-shot OOD detection setup, add-ons are added to the pre-trained model to better adapt it to the OOD detection task. ZOC [10] achieves zeto-shot OOD detection by training an image description generator on a large image captioning dataset [153], enabling the model to generate candidate unseen labels. It's essential to note that ZOC treats CLIP merely as a feature extractor and doesn't impart OOD detection capabilities to CLIP itself. The subsequent work CLIPN [99] empowers CLIP with the ability to say "no" by a "no-prompt" encoder. Other approaches alleviate the need for additional training and focus on enhancing the performance of post-hoc zero-shot OOD detection. A simple baseline of clip-based zero-shot OOD detection is to use the normalized text-image similarity as an OOD score. Ming et al. [98] further substitute the score with the maximum concept matching (MCM) score and offer a comprehensive theoretical explanation of this modification. Moreover, NegLabel [154], LAPT [102] and CLIPScope [96] enriches the text information through a predetermined corpus. Concretely, NegLabel selects negative words (labels) from the corpus, which are dissimilar to ID labels, to enhance OOD detection performance. Futhermore, CLIPScope uses those labels to revise the original score by applying Bayesian rules. In contrast, LAPT leverages text-to-image generation models or image retrieval models to obtain images corresponding to the given ID and Negative ID labels, followed by prompt tuning. It is noteworthy that ZOC, CLIPN, NegLabel and CLIPScope each devise unique OOD scores for integrating into their respective models. Conversely, the MCM score stands out for its generalizability across different models. The calculation of MCM is as follows:

Table 2. Performance evaluation of some competitive methods using the ImageNet-1K dataset as the ID dataset and iNaturalist, SUN, Places, and Textures as OOD datasets. The best result is emphasized in bold.

Scenario	Method	iNaturalist		SUN		Places		Textures		Average	
		AUROC ↑	FPR95 ↓	AUROC ↑	PFR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
Zero-shot	ZOC [10]	86.09	87.30	81.20	81.51	83.39	73.06	76.46	98.90	81.79	85.19
	MCM [98]	94.59	32.20	92.25	38.80	90.31	46.20	86.12	58.50	90.82	43.93
	CLIPN [99]	95.27	23.94	93.93	26.17	92.28	33.45	90.93	40.83	93.10	31.10
	NegLabel [154]	**99.49**	**1.91**	95.49	**20.53**	91.64	35.59	90.22	43.56	94.21	25.40
Few-shot	CoOp [158]	93.77	29.81	93.29	40.83	90.58	40.11	89.47	45.00	91.78	51.68
	LoCoOp [111]	96.86	16.05	95.07	23.44	91.98	32.87	90.19	42.28	93.52	28.66
	NegPrompt [112]	98.73	6.32	**95.55**	22.89	**93.34**	**27.60**	**91.60**	**35.21**	**94.81**	**23.01**

$$S_{MCM}(x';\mathcal{Y}_{in},T,\tau)=\max_i \frac{e^{s_i(x')/\tau}}{\sum_{j=1}^{K} e^{s_j(x')/\tau}}, \tag{6}$$

where x' is the test input image; $\mathcal{Y}_{in}=\{y_1,y_2,\ldots,y_K\}$ represents the label set; y_i denotes the i-th label in the label set; $T(t_i)$ corresponds to the concept vector for the text prompt t_i; $s_i(x')$ is the cosine similarity score between the image feature $I(x')$ and the concept vector $T(t_i)$, computed as $s_i(x')=\frac{I(x')\cdot T(t_i)}{\|I(x')\|\|T(t_i)\|}$; $e^{s_i(x')/\tau}$ is the exponentiated similarity score normalized by the temperature parameter τ; K is the number of labels in the label set; τ is the temperature parameter of the softmax function; and S_{MCM} is the maximum concept matching score. Some works have explored variations of zero-shot OOD detection or more challenging scenarios, such as detecting ID objects [155] or handling labels with noise [156], which will be discussed in the Section.7.

LLM-based. With the flourishing development of LLMs, new opportunities have emerged in the field of OOD detection with the application of LLMs. The strength of LLMs resides in their extensive world knowledge, enabling them to furnish comprehensive details about ID labels. Nevertheless, the issue of generating false or misleading information, known as hallucinations, poses a significant challenge when employing LLMs for OOD detection tasks. Existing methods [103, 104] are based on an observation: two categories of images may have very similar visual features but different properties in the semantic space. Dai et al. [103] introduce a consistency-based calibration method with an object detector to mitigate hallucinations and utilize LLMs' world knowledge for describing ID classes. However, Huang et al. [104] propose ODPC to generate peer-class based on ID labels and OOD samples within text and image modality. It should be noted that ODPC requires ID images for training MLP, indicating that it is not a zero-shot method. Regarding the OOD score, Dai et al. [103] adopt MSP, however, ODPC uses a KNN-based score. Later, a study [105] on LLMs was proposed to study three problems with LLMs doing OOD detection: the propensity of LLMs for near-OOD and far-OOD, the effect of different fine-tuning methods, and the choice of OOD score. Liu et al. [105] find that LLMs are natural far-OOD detectors, generative fine-tuning is better than discriminative fine-tuning, and cosine distance is sufficient as OOD score because the embedding space of LLMs exhibits anisotropy.

Remark. As mentioned above, "zero-shot" here refers to no exposure to ID images and only access to ID labels. However, Fort et al. [157] argue that in deployment scenarios, labels for OOD are either readily available. They propose a variant of the baseline approach that incorporates the names of OOD classes as candidate labels, resulting in improved performance. However, access to OOD labels is rare in real-life scenarios and seems less reasonable.

5.2 Few-shot OOD Detection Approaches

Overview. Given the large pre-trained model and a few ID data, we can adapt the model using the ID data and subsequently detect OOD test data. Zero-shot OOD detection does not necessitate any training images, making it suitable for scenarios with high-security requirements. However, it may face challenges related to domain gaps with ID downstream data, which can limit the performance of zero-shot methods. Therefore, there are many few-shot methods employed in OOD detection, and their effectiveness is often superior to that of zero-shot OOD detection.

Studies. The direct way of adapting a large pre-trained model with several ID samples is fine-tuning. Ming and Li [159] and Dong et al. [110] conduct studies on the impact of fine-tuning on OOD detection within the context of VLMs, with a special focus on CLIP. Ming and Li [159] pay more attention to the role of parameter-efficient fine-tuning (PEFT) methods and the OOD score. The MCM score, introduced by Ming et al. [98] as an innovative measure, alongside methods based on prompt learning for detecting OOD instances, is recognized for its effectiveness. Similarly, Dong et al. [110] conduct an extensive comparison between various fine-tuning approaches, such as PEFT and traditional methods, and discover that PEFT exhibits superior performance in detecting OOD instances, which echos with the conclusion of **?**]. Following this, Kim et al. [160] believe that Finetune Like You Pretrain (FLYP), a fine-tuning method, deserves attention due to its good performance on classification tasks. Specifically, FLYP is to mimic contrastive language-image pre-training as CLIP. After comparing the performance of FLYP and PEFT methods on zero-shot OOD detection, Kim et al. [160] find FLYP yields better OOD detection performance than PEFT. It should be noted that while Fort et al. [157] discuss "few-shot" OOD detection, their approach relies on a small number of outlier examples rather than ID samples, making it unsuitable for inclusion in this section.

Fine-tuning based. A prevalent method for fine-tuning CLIP using limited ID data involves prompt learning [161, 162], where the fine-tuning focuses on the prompt's contextual words, maintaining the pre-trained parameters unchanged. To distinguish between ID and OOD data, both LoCoOp [111] and ID-like Prompting Learning [109] are designed to enhance the learning of context vectors from the near-ID feature perspective. LoCoOp employs entropy maximization to distance ID-irrelevant local features (such as the backgrounds in ID images) from the textual embeddings of ID classes. Similarly, ID-like Prompting Learning generates ID-like data—outliers within the vicinity of ID samples—to refine the context vector. Additionally, it incorporates a diversity loss to enhance the variance among the sampled OOD data. However, recent work EOK-Prompt [107] and GalLop [108] argues that optimizing only a prompt using numerous features from an image wastes valuable information. Instead, EOK proposes an orthogonal method to optimize the novel local prompt, making better use of the limited image data, while GalLop optimizes global prompts and local prompts by leveraging global and local features, respectively. Furthermore, based on the mentioned studies, DSFG [110] posits that traditional fine-tuning approaches applied to CLIP may inadvertently lead to a loss of critical OOD knowledge. To address this issue, DSFG adopts a strategy to retain broad OOD knowledge by merging the original features with those modified through fine-tuning, followed by the training of a classifier. DSFG's plug-and-play capability makes it seamlessly compatible with all fine-tuning-based approaches, enhancing its practicality and value. The aforementioned few-shot tuning methods inevitably suffer from overfitting on limited-shot instances. Chen et al. [163] proposed a training-free few-shot OOD detection method, Dual-Adapter[163], which constructs two types of adapters by extracting positive and negative features to aid in OOD detection. The currently mentioned few-shot methods lack OOD supervision, so CLIP-OS [106] tries to find such supervision and also achieves stunning results

Meta-learning based. Meta-learning aims to devise a learning approach that enables rapid adaptation to new challenges [164, 165]. OOD-MAML [115] adapt model-agnostic meta-learning (MAML) for few-shot OOD detection. It

generates OOD samples and incorporates them along with ID data for the adapted N-way K-shot task, which is divided into N sub-tasks, each focusing on K-shot OOD detection. The decision on whether test data is OOD is based on the outcomes of these fast and simple N sub-tasks. In contrast, HyperMix [116] advocates for employing a hypernetwork-based method to enhance sample augmentation without the necessity for extra outliers. This is because classes not included in a specific meta-training task can act as OOD samples.

5.3 Full-shot OOD Detection Approaches

Overview. This setup is generally less realistic than the first two (zero-shot and few-shot). However, we list them separately to ensure a comprehensive review of existing methods across the spectrum. Given the full set of ID data and corresponding labels, VLMs can enhance OOD detection significantly by fine-tuning. Moreover, a novel task called "PT-OOD" detection is introduced.

Fine-tuning based. With access to the complete dataset, then more data can be used to fine-tune the large pre-trained model or the data can be used to better simulate the ID distribution, facilitating the differentiation of OOD data. NPOS [29] proposes a non-parametric outlier synthesis technique to distinguish ID and OOD data by fine-tuning CLIP with complete ID data. In contrast, TOE [118], while also using CE loss to constrain the model during fine-tuning, builds on the ideas of OE by focusing on textual outliers within the CLIP framework to control the model's recognition capabilities, which differs significantly from directly using OOD images.

PT-OOD Detection. "PT-OO" samples are OOD samples with overlap in pretraining data. After investigating and elucidating the effects of various pre-training methodologies (supervised, self-supervised) on PT-OOD detection, Miyai et al. [117] observe the low linear separability in feature space significantly degrades the PT-OOD detection performance. They suggest using distinctive features for each instance to distinguish between ID and OOD samples.

6 EVALUATION AND APPLICATION

6.1 Evaluation metrics

In the vast majority of OOD detection tasks in the visual domain, the following evaluation metrics are commonly used:

AUROC (Area Under the Receiver Operating Characteristic curve). This metric quantifies the likelihood that a classifier will assign higher scores to ID samples compared to OOD samples. An elevated AUROC value is indicative of superior model performance, signifying an enhanced ability to distinguish between ID and OOD instances. Consequently, a higher value is desirable.

AUPR (Area under the Precision-Recall curve). This metric is pertinent when the ID class is considered the positive class and is particularly valuable in the context of imbalanced class distributions. It assesses the balance between precision and recall, with a higher AUPR value indicating superior model performance. Therefore, a greater AUPR value is desirable.

FPR@95 (False Positive Rate at 95% True Positive Rate). This metric delineates the false positive rate (FPR) at the juncture where the true positive rate (TPR) reaches 95%. It essentially gauges the proportion of OOD samples erroneously identified as ID, thus providing insight into the model's propensity for false alarms at a high sensitivity threshold. A reduced FPR@95% TPR is indicative of a model's enhanced specificity in correctly flagging OOD samples while maintaining high sensitivity towards ID samples. Therefore, a lower value is desirable.

These metrics provide comprehensive insights into the OOD detection performance, considering different aspects such as the ability to distinguish between ID and OOD samples, handling imbalanced class sensitivity towards ID samples, and controlling the false positive rate at a specific true positive rate threshold.

6.2 Experimental Protocols

In the traditional experimental protocol for OOD detection, test data is exclusively classified as either ID or OOD. However, as the field has advanced, there is now a more nuanced distinction between OOD and ID data, which has led to variations in the evaluation process.

Subsequently, OOD data is categorized into near-OOD and far-OOD based on the degree of covariate shift from ID data. This categorization corresponds to dividing OOD detection tasks into near-OOD and far-OOD detection. It is evident the near-OOD detection task is more challenging, however, numerous methods [98, 109, 157] have demonstrated excellent performance in this area.

Recently, Yang et al. [33], Bai et al. [166] propose that we should consider cases where covariate shift occurs in ID data, which is not taken into account previously. This is crucial to prevent the loss of model generalization. The samples mentioned earlier are termed as cs-ID data, an abbreviation for "covariate shift ID" data. Consequently, a new experimental protocol has been explored, called full-spectrum OOD detection. During the testing phase, the model is expected to identify near-OOD and far-OOD instances. Additionally, it should refuse to provide predictions for the OOD data and accurately predict ID and cs-ID data.

6.3 Application

6.3.1 Computer Vision. Most of the efforts in OOD detection have been devoted to the field of computer vision, we list extensive vision-related tasks as follows:

- **Image Classification.** The majority of tasks discussed in this paper focus on OOD detection within the realm of image classification. In such task scenarios, commonly utilized ID datasets include MNIST [172], CIFAR-10 [173], CIFAR-100 [173], and ImageNet-1K [174]. Various OOD datasets are constructed accordingly for evaluating different methods [175], with the most frequently employed datasets being iNaturalist [176], SUN [177], Places [178] and Textures [179]. Additionally, full-spectrum OOD is typically evaluated on three benchmarks: DIGITS, OBJECTS, and COVID, as proposed by Yang et al. [33].
- **Semantic Segmentation.** Recent works [90] have started delving into the dense OOD detection task, also known as anomaly segmentation. The datasets used for evaluation include the Cityscapes dataset [180], the Road Anomaly dataset [181], and the recently developed SOOD-ImageNet [182].
- **Object Detection.** The application of methods related to OOD detection in the field of object detection is relatively nascent, with only a few studies exploring this area [34]. Evaluations are commonly performed using datasets like PASCAL-VOC [183] and Berkeley DeepDrive-100K [184].
- **Autonomous Driving.** Autonomous driving has long been a crucial practical application of OOD Detection. Recently, Mao et al. [167] utilized the CARLA [185] system to simulate and evaluate the performance of OOD Detection in autonomous driving scenarios.
- **Medical Image Analysis.** In the field of medical image analysis, OOD detection is crucial. Depending on the specific category of medical image, OOD detection employs various datasets including CIFAR-10 and Kvasir-Capsul [186].

Table 3. Summary of Datasets. The CARLA System is a simulation platform designed for evaluating OOD Detection in the field of autonomous driving, hence its entire row is filled with dashes. Other "-" symbols represent numbers that vary depending on usage. More detailed descriptions can be found in the code repo.

TASK	Dataset Name	Data Type	# Classes	# Samples	Papers
Image Classification	CIFAR-10	Images	10	60,000	[10, 99]
	CIFAR-100	Images	100	60,000	[10, 99]
	MNIST	Images	10	70,000	[2, 81]
	ImageNet-1K	Images	1,000	1,431,167	[98, 99]
	iNaturalist	Images	5,089	675,170	[98, 99]
	SUN	Images	397	108,754	[98, 99]
	Places	Images	>205	>2,500,000	[98, 99]
	Textures	Images	47	5,640	[98, 99]
Semantic Segmentation	Cityscapes	Images	-	25,000	[90]
	Road Anomaly Dataset	Images	-	100	[90]
Object Detection	PASCAL VOC	Images	20	2,913	[34, 97]
	BBD100K	Video->Image	Variable	100,000	[34, 97]
Autonomous Driving	CARLA System	-	-	-	[167]
Medical Image Analysis	Kvasir-Capsul	Images		4,741,621	[168]
Text Category	News Category	Text	-	210,000	[169]
	SST-2	Text	-	215,154	[169]
Intent Detection	CLINC150	Text	150	22,500	[170]
	Banking	Text	77	13,083	[170]
	StackOverflow	Text	20	20,000	[170]
Audio	MSCW	Audio	-	>23400000	[171]
	Vocalsound	Audio	-	21,024	[171]
Graph data	TU	Graph data	-	Variable	[94]
	OGB	Graph data	-	Variable	[94]

OOD detection has significant applications across various fields such as human action recognition [187] and solar image analysis [188, 189]. For further details, please refer to the accompanying code repository due to space limitations. The same applies to the subsequent paragraphs.

6.3.2 Natural Language Processing. OOD detection is also explored in various tasks across numerous Natural Language Processing applications. The two most common applications are as follows:

- **Intent Detection.** Intent Detection is a significant application of OOD detection in NLP. The datasets used for evaluation include CLINC150 [190], Banking [191], StackOverflow [192], among others.
- **Text Catagory.** In text category OOD detection applications, datasets like News Category [193] and SST-2 [194] are commonly utilized to form ID/OOD pairs and assess the models' detection capabilities.

6.3.3 Beyond Computer Vision and Natural Language Processing. In addition to the two data modalities mentioned above, OOD detection still has many important applications across various types of data.

- **Audio data.** In audio OOD detection, MSCW (Micro-EN) [195] and Vocalsound [196] are usually used as ID datasets, and they also act as OOD for each other.
- **Graph data.** Recent studies have proposed various approaches for OOD detection in graph data. The existing graph-level OOD detection benchmark comprises a total of 10 pairs of datasets from TU [197] and OGB datasets [198], which has been widely utilized.
- **Reinforcement learning.** Currently, there is a rising trend of integrating OOD detection with reinforcement learning [199] to bolster model robustness. Mohammed and Valdenegro-Toro [200] provide a benchmark and explore methods for crafting tailored reinforcement learning environments that can generate OOD data.

7 EMERGING TRENDS AND OPEN CHALLENGES

Despite rapid advancements in OOD detection, numerous emerging trends and less-explored challenges remain. In this section, we explore emerging trends and open challenges from three distinct perspectives: methodologies, scenarios, and applications.

7.1 Better methodologies of OOD detection

Meta-Learning Adaptation. Faced with the quick adaptation challenge in test-time OOD detection, mete-learning algorithms, which provide a learning-to-learn paradigm to efficiently adapt the model to new test data, may be the solution. Additionally, in addressing the dauntingly vast sample space of potential outliers inherent in training-driven OOD detection, improved sampling methods could offer a pathway to the efficient utilization of outliers [52].

Theoretically-Driven Score Designing. In the traditional single-modal regime, numerous post-hoc scores have been carefully designed for OOD detection, effectively reducing training costs. However, as the field has progressed into the multi-modal domain, there is an increasing demand for theoretically-driven score designs. MCM [98] is a notable example, but it is merely an extension of Softmax and does not deeply explore the relationships between text and image. Therefore, more advanced scores are needed.

7.2 More practical scenarios of OOD detection

Under the current trend, there is a growing need for the emergence of more practical scenarios, driven by the limitations of existing impractical restrictions.

Quick Test-Time Adaptation. The advent of test-time scenarios like CAOOD [144] significantly advances the application of OOD detection in real-world contexts, promising improved reliability and adaptability.

Multi-Modal Detection. Exploring multi-modal OOD detection enhances our grasp of data dynamics and boosts model efficacy across varied sensory inputs. Additionally, the forthcoming integration of LLMs into multi-modal OOD

detection stands to transform the field, marrying advanced linguistic analysis with other data types to forge more effective and versatile detection systems for complex applications.

Open-Vocabulary Scenario. Existing LPM-based methods assume that ID categories are accessible. However, this assumption does not hold in open-vocabulary scenarios. Li et al. [112] addresses this by learning migratable negative prompts. It uses a subset of ID labels for prompt learning, and these learned prompts can be applied to other unseen labels. OOD detection in open-vocabulary settings further relaxes the requirement for prior access to ID knowledge, paving the way for the development of more generalized models.

Noise Settings. While previous research has largely focused on standard clean label settings, few studies have explored the other side of the coin: noisy label scenarios. Humblot-Renaux et al. [201] investigate practical scenarios and examine the impact of varying levels of noise on OOD detection. Then they provide several key takeaways for improving OOD detection in noisy label environments.

7.3 New applications of OOD detection

Additional Modalities. While OOD detection has made strides across diverse sectors, its potential in speech and physiological signal analysis remains largely untapped. In particular, OOD issues are prevalent in emotion-related physiological signals due to the variability in subjects' emotional states. Therefore, this represents a promising area for further exploration. Some of the current methods [202] also provide benchmarks for reference.

Human-in-the-loop Application. A pivotal future direction, proposed by Vishwakarma et al. [203], is the integration of human insight into the detection process. This human-in-the-loop approach, particularly in high-stakes decision-making, will be vital for enhancing the accuracy and responsiveness of OOD detection systems, merging human intuition with algorithmic precision.

Web Image Scraping. To automate the process of scraping images from the Internet, Miyai et al. [155] propose a zero-shot ID detection task, a concept derived from zero-shot OOD detection. An image will be classified as an ID image if it contains ID objects; it will only be considered an OOD image if it lacks any ID objects. It presents a novel perspective on the application of OOD detection and merits further exploration

8 CONCLUSION

OOD detection is critical for trustworthy machine learning. In this paper, we provide a comprehensive review of recent advances in OOD detection, focusing for the first time on the problem scenario perspective: training-driven, training-agnostic, and large pre-trained model-based OOD detection. We also summarize extensively used evaluation metrics, experimental protocols, and diverse applications. We believe that our novel taxonomy of existing papers and extensive discussion of emerging trends will contribute to a better understanding of the current state of research, assist practitioners in selecting suitable approaches, and inspire new research hotspots.