_2_motivation:_the_imperative_for_trustworthy_ai _2_post-hoc_confidence-based_detection _3_feature-space_distance-based_methods _2_reconstruction_autoencoders:_advancements__and__limitations _3_energy-based_models_for_ood_detection _2_learning_robust__and__separable_feature_representations _3_gradient-based__and__neuron-level_analysis _2_ood_in_specialized_learning_settings _3_leveraging_pre-trained_foundation_models _2_ood_for_autonomous_systems__and_physical_systems _3_ood_in_cybersecurity__and__anomaly_detection _4_practical_deployment_considerations__and__human-in-the-loop _2_certifiable_ood_detection:_provable_guarantees__and__conformal_prediction _3_st_and_ar _2_open_challenges__and__future_research_avenues _3_ethical_considerations__and__societal_impact

A Comprehensive Literature Review with Self-Reflection

Literature Review

October 7, 2025

Abstract

This literature review provides a comprehensive analysis of recent research in the field. The review synthesizes findings from 186 research papers, identifying key themes, methodological approaches, and future research directions.

Contents

1 Introduction to Out-of-Distribution Detection

1.1 Defining Out-of-Distribution Data and Distribution Shifts

The robust deployment of machine learning models in real-world, open-world environments critically hinges on their ability to recognize when input data deviates from the distribution they were trained on. This fundamental challenge is addressed by Out-of-Distribution (OOD) detection, a field dedicated to distinguishing In-Distribution (ID) data, which models are designed to process, from novel, unfamiliar OOD data. Initially, OOD was often conceptualized as a straightforward "semantic shift," implying entirely new classes or concepts unseen during training. However, this definition has evolved significantly to encompass a more complex spectrum of distribution shifts that profoundly challenge model generalization.

Early work often struggled to consistently categorize various forms of data shifts. (3) critically addressed this by introducing the "Full-Spectrum OOD (FS-OOD)" problem, explicitly distinguishing between **semantic shift** (novel classes) and **covariate shift** (changes in input appearance or style, such as lighting or viewpoint, while retaining the same semantic class). Their proposed SEM score function, which disentangles semantic and non-semantic features, aimed to detect true semantic novelty while remaining robust to covariate variations, highlighting the necessity of a nuanced approach beyond simple binary classification. Further complicating this, (8) introduced the concept of "spurious OOD," demonstrating that models can make overconfident predictions on OOD inputs that share spurious correlations with ID data, even if they lack the essential invariant features. This revealed a deeper vulnerability where models exploit non-causal features, making detection particularly challenging and underscoring the inherent ambiguity in OOD boundaries.

The need for more rigorous evaluation and clearer definitions led to significant advancements in benchmarking. (54) meticulously curated `ImageNet-OOD`, a dataset designed to isolate pure semantic shift by removing ID contamination, semantic ambiguities, and unintended covariate shifts prevalent in prior benchmarks. This effort emphasized the

difficulty in creating truly clean OOD definitions and highlighted how existing methods often inadvertently detected covariate shifts rather than genuine semantic novelty. Building on this, (56) proposed a "Model-Specific Out-of-Distribution (MS-OOD)" framework, which redefined OOD not solely by data properties but by whether a *deployed model* could correctly classify an example. This unified the detection of semantic shift, covariate shift (when misclassified), and even misclassified ID examples under a single, performance-driven ground truth, providing a more practical and holistic perspective. The "Sorites Paradox" of OOD, where the degree of shift is continuous rather than binary, was addressed by (178). They introduced the "Incremental Shift OOD (IS-OOD)" benchmark and the LAID method, which leverages CLIP to decompose image features into distinct semantic and covariate components, allowing for a continuous measurement of shift levels. This moved the field towards understanding OOD as a spectrum rather than a discrete boundary. (140) further dissected OOD detection and open-set recognition, providing a critical analysis of methods and benchmarks, and emphasizing the need for evaluation protocols that disentangle semantic and covariate shifts, especially at scale where methods like Outlier Exposure struggle due to the difficulty of acquiring representative auxiliary OOD data.

Beyond visual data, the definition and challenges of OOD extend to other modalities. (6) pioneered unsupervised OOD detection for graph-structured data with GOOD-D, addressing the unique topological and feature-based shifts in graphs. (100) scaled OOD detection to multimodal settings, introducing the `MultiOOD` benchmark and the Agree-to-Disagree (A2D) algorithm to leverage complementary information across modalities (e.g., video, audio, optical flow) and identify `Modality Prediction Discrepancy` as an OOD signal. For natural language processing, (24) provided a comprehensive survey, highlighting the distinct challenges of discrete input spaces and contextual semantic shifts. In generative language models, (107) tackled OOD detection in mathematical reasoning, identifying "pattern collapse" in output spaces and proposing a "Trajectory Volatility Score" based on dynamic embedding changes, demonstrating how domain-specific phenomena necessitate specialized OOD definitions. Furthermore, theoretical investigations

have deepened our understanding of OOD learnability. (126) explored the PAC learnability of OOD detection, proving that it is not universally learnable and depends critically on the characteristics of the data distributions and hypothesis spaces. (121) provided theoretical conditions for *when and how* in-distribution labels help OOD detection, particularly for "near OOD" scenarios. Finally, (51) offered a theoretical explanation for the efficacy of feature norms in OOD detection, linking it to hidden classifier confidence and proposing a "Negative-Aware Norm" (NAN) that accounts for both activation and deactivation tendencies of neurons, providing a deeper insight into the internal mechanisms that differentiate ID from OOD.

In conclusion, the definition of OOD data and distribution shifts has evolved from a simplistic notion of novel classes to a multifaceted concept encompassing semantic, covariate, and spurious shifts, often viewed as a continuous spectrum rather than a hard boundary. The field now grapples with model-specific interpretations, multimodal challenges, and fundamental questions about learnability, necessitating robust detection mechanisms that are sensitive to diverse forms of unfamiliarity while being resilient to expected variations. The inherent ambiguity in precisely delineating OOD boundaries remains a central, ongoing challenge in the field.

1.2 Motivation: The Imperative for Trustworthy AI

The increasing integration of artificial intelligence (AI) systems into critical societal infrastructures and high-stakes applications necessitates an unwavering commitment to trustworthiness, reliability, and safety. A fundamental challenge that directly undermines this trust is the inherent overconfidence of deep learning models when confronted with inputs that deviate significantly from their training distribution, commonly referred to as Out-of-Distribution (OOD) data. This section articulates the compelling and urgent reasons behind the escalating importance of OOD detection in modern AI, emphasizing its role as an indispensable component for building truly trustworthy, robust, and safe artificial intelligence.

In numerous safety-critical domains, the consequences of unchecked model overconfi-

dence on OOD data can be catastrophic, leading to unreliable decisions, system failures, and potentially severe harm. For instance, in autonomous driving, a vehicle's perception system misinterpreting an anomalous road condition, an unfamiliar object, or an unusual weather pattern as a familiar in-distribution (ID) input can lead to dangerous maneuvers or accidents (20; 17). Similarly, in medical diagnosis, an AI system providing a highly confident but incorrect diagnosis for a rare or unseen patient condition, or misinterpreting an anomalous medical image, could have dire implications for patient well-being (16; 17). The deployment of deep reinforcement learning (RL) agents in real-world control systems also faces this challenge, where agents trained in simulated environments may encounter novel states in the physical world and fail silently without signaling uncertainty, posing significant safety risks (147). This imperative for trustworthy AI demands that these systems not only perform well on familiar data but also express meaningful and calibrated uncertainty when encountering novel, unfamiliar, or anomalous inputs (4; 12; 141).

The core problem stems from the "closed-world" assumption under which most deep learning models are traditionally trained. This assumption posits that test data will be drawn from the same statistical distribution as the training data (32). However, this premise rarely holds true in complex, dynamic, and open-world environments where unforeseen circumstances, sensor noise, adversarial attacks, or simply novel data points are inevitable (4; 11; 12; 29; 89; 118). When this closed-world assumption is violated, conventional models often produce high-confidence, yet incorrect, predictions for OOD samples (5; 25; 27; 42; 141). This unwarranted overconfidence is a critical vulnerability that OOD detection aims to mitigate, providing a crucial safety mechanism to prevent models from making decisions outside their learned competence.

Furthermore, the very nature of OOD data can be complex and multifaceted, posing additional challenges to reliable detection. It is not always a simple binary distinction between "known" and "unknown." For instance, models can learn spurious correlations from their training data, leading them to confidently classify OOD inputs that share these irrelevant features as in-distribution, making such "spurious OOD" particularly difficult to detect (8). This highlights that a robust OOD detector must not only identify

entirely novel semantic concepts but also be resilient to subtle shifts or misleading cues. Moreover, the definition of what constitutes "OOD" can even be model-specific, depending on whether a particular input leads to a misclassification for a given deployed model, rather than a universal distributional shift (56). These nuances underscore the need for sophisticated and context-aware OOD detection mechanisms.

The practical deployment of AI systems further amplifies the need for robust OOD detection. Beyond theoretical performance, real-world systems require OOD detectors that are not only accurate but also provide reliable guarantees and manage false positives effectively. High false positive rates (FPR), where legitimate in-distribution samples are incorrectly flagged as OOD, can lead to user frustration, unnecessary human intervention, and a breakdown of trust in the system (118). Therefore, the development of OOD detection is intrinsically linked to the broader goal of building AI systems that are transparent about their limitations, can abstain from making potentially harmful decisions when faced with unfamiliar situations, and can operate predictably and safely in dynamic environments.

In summary, the motivation for robust OOD detection is deeply rooted in the urgent need to transition AI from research curiosities to reliably deployed systems that operate safely and responsibly in the real world. It is not merely about identifying novelty but about ensuring that AI systems are aware of their limitations, can express appropriate uncertainty, and can defer to human oversight or alternative safe actions when confronted with inputs beyond their learned experience. OOD detection, therefore, stands as a pivotal step towards enabling the responsible and reliable deployment of AI in complex, open-world environments, where unforeseen circumstances are not exceptions but inevitable realities. Continued research in this area is essential to bridge the gap between theoretical capabilities and the practical demands of trustworthy AI.

2 Foundational Concepts and Early Post-Hoc Methods

2.1 Uncertainty Quantification in Neural Networks

Conventional neural networks, while achieving remarkable performance in complex classification tasks, are fundamentally designed to assign inputs to a fixed set of predefined classes rather than to explicitly quantify the uncertainty inherent in their predictions. Nevertheless, these classifiers offer implicit signals of confidence primarily through softmax probabilities and the entropy of the predicted class distribution. These metrics serve as initial, easily accessible indicators of a model's belief. A critical understanding of their foundational role and, more importantly, their profound limitations is indispensable for appreciating the subsequent evolution of Out-of-Distribution (OOD) detection methodologies that strive for truly calibrated uncertainty estimates.

A seminal contribution by (**?**) formalized the use of Maximum Softmax Probability (MSP) as a straightforward baseline for identifying both misclassified in-distribution (ID) and OOD examples. This work, alongside others, critically exposed a pervasive and dangerous flaw: neural networks frequently exhibit severe overconfidence, assigning high softmax probabilities to OOD inputs. This leads to erroneous high-confidence predictions that are fundamentally unreliable, as the model confidently asserts an input belongs to a known class despite having never encountered anything similar during training. The root cause of this overconfidence lies in the very design of the softmax function. As (**?**) elucidates, softmax is inherently a normalized probability distribution over a fixed set of *known* classes. It is optimized to express the model's certainty about which of the *trained* categories an input belongs to (reflecting *aleatoric uncertainty* due to inherent data noise), but it is ill-equipped to capture *epistemic uncertainty*—the model's lack of knowledge or confidence when confronted with inputs far removed from its training distribution. Consequently, an OOD input, by definition, falls outside the model's learned domain, yet the softmax mechanism forces it into one of the known categories, often with high confidence, simply by finding the "closest" match within its learned manifold.

Beyond this conceptual mismatch, modern deep neural networks are frequently poorly

calibrated (**?**). This means their predicted probabilities do not accurately reflect the true likelihood of correctness, exacerbating the overconfidence problem, particularly for OOD inputs. The architectural choices prevalent in deep learning, such as increased depth, ReLU activations, and optimization for accuracy rather than calibration, contribute to this miscalibration. When a model's confidence scores are unreliable even for ID data, their utility for discerning OOD samples becomes severely compromised. Furthermore, the issue of overconfidence is significantly compounded by the model's reliance on spurious correlations present in the training data. As (8) rigorously demonstrated, models trained on datasets containing statistically informative but non-causal features tend to exploit these shortcuts. When an OOD input shares these spurious features with ID data, the model can confidently assign a high softmax probability, even if the input lacks the invariant, semantic features crucial for correct classification. This reliance on misleading environmental cues makes distinguishing spurious OOD samples from ID data inherently challenging, as the model's confidence is rooted in a superficial correlation rather than true semantic understanding (8).

Empirical and theoretical studies consistently underscore the inadequacy of raw softmax and entropy scores for robust OOD detection. (38) provided extensive evidence that simple prediction-based methods like MSP and entropy are reliably outperformed by methods leveraging learned intermediate representations (embeddings). Their work challenged the notion of MSP as a universally strong baseline, demonstrating that while it might offer some rudimentary signal, it often falls short compared to approaches that analyze the internal feature space, which are better equipped to capture deviations from the ID manifold. This empirical observation is theoretically grounded by (104), who critically analyzed logit-based methods, including those derived from softmax. They explained that these methods are often not directly proportional to true data density. This fundamental disconnect implies that even when a model's logits are high, the resulting softmax probability does not necessarily reflect a high likelihood under the true in-distribution data manifold, leading to suboptimal OOD detection performance. Similarly, (56)'s comprehensive evaluation, while introducing a model-specific perspective, implicitly highlights

the context-dependent and often inconsistent performance of MSP across different types of OOD shifts (e.g., semantic vs. covariate) and misclassifications, reinforcing its limitations as a standalone, universally reliable uncertainty measure.

In summary, while softmax probabilities and entropy offer initial, easily accessible indicators of a neural network's confidence, their inherent limitations are profound and multifaceted. These include their inability to capture epistemic uncertainty due to their closed-set design, the pervasive problem of miscalibration in modern deep networks, and their vulnerability to spurious correlations in training data. These shortcomings collectively render raw output-based uncertainty estimates unreliable for robust OOD detection, particularly in safety-critical applications where silent failures can have severe consequences. This fundamental inadequacy necessitates the development of more sophisticated OOD detection methodologies that move beyond simple output scores, paving the way for the advanced post-hoc, feature-space, generative, and training-time strategies discussed in subsequent sections of this review.

2.2 Post-Hoc Confidence-Based Detection

The development of robust out-of-distribution (OOD) detection methods is paramount for ensuring the reliability and safety of machine learning systems in real-world applications. Early and highly influential research in this domain focused on leveraging and refining confidence scores derived from pre-trained discriminative classifiers. These "post-hoc" methods are particularly attractive due to their efficiency, as they do not require any model retraining or modification of the original classification objective, thereby minimizing computational overhead and preserving the model's primary task performance. This section details the evolution of such confidence-based approaches, from foundational baselines to sophisticated enhancements.

A seminal contribution to this field was the introduction of Maximum Softmax Probability (MSP) as a baseline for OOD detection by (**?**). This straightforward yet surprisingly effective method operates on the premise that a well-trained classifier should assign a high maximum softmax probability to in-distribution (ID) samples, reflecting strong

confidence in its classification. Conversely, OOD samples, which do not align with any learned class, are expected to yield lower maximum probabilities. Despite its simplicity, MSP established a crucial benchmark, demonstrating that standard neural network outputs inherently contain valuable uncertainty signals. The work also played a pivotal role in standardizing evaluation protocols and datasets, fostering more rigorous comparisons across diverse OOD detection techniques. However, a significant limitation of MSP is the pervasive problem of neural network overconfidence on OOD inputs (**?** **?**). Models can often assign spuriously high confidence to novel, unseen data, especially for "near OOD" examples that share superficial similarities with ID data, leading to suboptimal discrimination and false negatives. This fundamental challenge highlights that raw softmax probabilities, while indicative, are not always reliable estimators of true data density or typicality (104).

Building upon the insights from MSP, (**?**) introduced Out-of-Distribution Detector for Neural Networks (ODIN), a method designed to significantly amplify the distinction between ID and OOD samples without requiring any model retraining. ODIN introduced two key innovations. First, it incorporated temperature scaling, a technique originally used for model calibration, which smooths the softmax distribution by dividing the logits by a temperature parameter T. This adjustment makes the confidence scores less extreme and often more discriminative for OOD detection by re-calibrating the output probabilities. Second, and more crucially, ODIN proposed a small, carefully crafted input perturbation. This perturbation is calculated to push the input towards the direction that maximizes the softmax probability for the predicted class. For ID samples, this makes them "more ID-like" in the model's perception, increasing their confidence. For OOD samples, which lack a strong alignment with any ID class, this perturbation often fails to significantly boost confidence or may even push them towards lower confidence, thereby increasing the separation in confidence scores between ID and OOD data. By combining these two simple yet powerful post-hoc techniques, ODIN substantially boosted OOD detection performance over MSP, establishing a strong, efficient baseline.

Further refining the ODIN paradigm, (**?**) proposed Generalized ODIN (G-ODIN),

which aimed to improve robustness by addressing a potential weakness in ODIN's perturbation strategy. While ODIN perturbs inputs to maximize confidence for the *predicted* class, G-ODIN considers a broader context. Instead of relying on a single predicted class, G-ODIN perturbs the input towards the direction that minimizes the maximum softmax probability across *all* ID classes. This approach makes the OOD score more robust by ensuring that an OOD sample is not mistakenly pushed to high confidence for an incorrect ID class. By considering the full set of ID classes during perturbation, G-ODIN can achieve better discrimination, especially when OOD samples might be ambiguous or share features with multiple ID categories.

The concept of temperature scaling, a cornerstone of ODIN, has continued to evolve. Recognizing that a fixed global temperature might not be optimal for all samples, (138) introduced Adaptive Temperature Scaling (ATS). ATS proposes dynamically calculating a *sample-specific* temperature value based on activations from intermediate layers of the neural network. By fusing this sample-specific adjustment with class-dependent logits, ATS captures additional statistical information that might otherwise be lost in the feature extraction process. This dynamic approach leads to a more robust and powerful OOD detection method, demonstrating that even subtle refinements to temperature scaling can significantly enhance the performance and robustness of existing logit-based OOD detection techniques.

While highly effective and efficient, these confidence-based methods fundamentally rely on the assumption that the classifier's output space (logits or softmax probabilities) can reliably distinguish between ID and OOD data. However, as highlighted by (104), raw logit-based scores, including energy scores (which are closely related to logits), often make implicit assumptions about the underlying data distribution, such as constant partition functions across classes, or that softmax probabilities directly represent true data densities. These assumptions are not always accurate, limiting the theoretical grounding and empirical robustness of simpler confidence scores. To address this, (104) proposed ConjNorm, a novel theoretical framework grounded in Bregman divergence, which unifies density function design for OOD detection within the exponential family of distributions.

By devising an unbiased and analytically tractable estimator for the partition function using importance sampling, ConjNorm offers a more principled and flexible approach to density estimation for OOD scoring, moving beyond restrictive distributional assumptions and leading to superior empirical performance. This work signifies a critical advancement towards more theoretically robust confidence-based OOD measures.

In summary, post-hoc confidence-based detection methods, starting from the simplicity of MSP and progressing through the innovative enhancements of ODIN, G-ODIN, and ATS, have laid a robust foundation for the field. They collectively demonstrated that significant improvements in uncertainty quantification could be achieved by cleverly leveraging and refining the outputs of existing pre-trained classifiers without the need for costly retraining. However, their inherent reliance on the classifier's output space and the potential for overconfidence or inaccurate density estimation remain persistent challenges. These limitations motivate the exploration of richer, intermediate representations and more theoretically grounded approaches to OOD detection, which are discussed in subsequent sections.

2.3 Feature-Space Distance-Based Methods

Early efforts in out-of-distribution (OOD) detection quickly recognized the inherent limitations of relying solely on a neural network's final output probabilities, such as Maximum Softmax Probability (MSP). While simple, MSP often fails to capture the true uncertainty for samples significantly deviating from the in-distribution (ID) manifold, frequently exhibiting overconfidence on novel inputs. This critical observation spurred a shift towards leveraging the internal feature representations of neural networks, driven by the hypothesis that OOD samples would manifest as distinct patterns or lie significantly distant from the ID data within these learned embedding spaces.

A foundational exploration into using internal representations for OOD detection was presented by (**?**). While this work primarily established MSP as a baseline, it also investigated the efficacy of Mahalanobis distance computed on features from intermediate layers. By modeling the distribution of ID features for each class as a simple Gaussian,

OOD samples could be identified as those with a large Mahalanobis distance to all ID class centroids. This early insight highlighted the potential of feature-level analysis to provide more robust OOD scores than simple output probabilities, laying crucial groundwork.

Building upon this concept, (**?**) introduced a more comprehensive framework that explicitly leverages Mahalanobis distance in the feature space for OOD detection. This method models the in-distribution feature representations for each class using class-conditional Gaussian distributions, where the mean and covariance are estimated from the training data. An input is then classified as OOD if its Mahalanobis distance to all ID class centroids is sufficiently large. Crucially, (**?**) also proposed using generative adversarial networks (GANs) to regularize the feature space during training. By training a GAN to generate OOD samples and then using these to push OOD representations away from ID clusters, the feature space becomes more discriminative, enhancing the separation between ID and OOD samples and making the Mahalanobis distance a more effective OOD score. This approach demonstrated a significant advancement by actively shaping the feature space to be more amenable for OOD discrimination, moving beyond merely observing existing features.

Despite its principled nature, the effectiveness of Mahalanobis distance-based methods can be limited by the strong assumption of Gaussianity for ID features and the quality of the learned features, particularly in high-dimensional spaces where the "curse of dimensionality" can render distance metrics less meaningful (53). This motivated further research into refining the feature space and the distance calculations themselves. For instance, (65) conducted an in-depth analysis of Mahalanobis distance for medical imaging OOD detection, challenging the common assumption of a single optimal layer for detection. They empirically demonstrated that the optimal network depth for OOD detection is highly dependent on the specific OOD pattern and proposed a Multi-branch Mahalanobis (MBM) framework. MBM employs multiple OOD detectors operating at different depths of the network, each combining normalized Mahalanobis scores from its constituent modules, significantly enhancing robustness by capturing diverse OOD signals across the feature hierarchy.

To mitigate the curse of dimensionality and improve feature space utility, researchers have explored learning more structured and compact representations. (7) proposed training deep neural networks to embed ID samples onto a union of 1-dimensional subspaces. This compact representation ensures that OOD samples are less likely to occupy the same region as known classes, and robust representatives (singular vectors) can be used for distance calculations, thereby simplifying OOD detection. Similarly, (53) introduced Subspace Nearest Neighbor (SNN), a framework that regularizes the model and its feature representation by leveraging the most relevant subset of dimensions. This subspace learning yields highly distinguishable distance measures between ID and OOD data, demonstrating significant improvements over previous distance-based methods by making the distances more robust to high-dimensional noise. Extending this, (162) proposed a novel "tangent distance" that explicitly accounts for the data structure by mapping high-dimensional features to the manifold of ID samples. This method computes the Euclidean distance between samples and the nearest submanifold space (a linear approximation of the local region on the manifold), providing a more meaningful distance measure that is less sensitive to the curse of dimensionality.

Beyond Mahalanobis and its direct refinements, other distance-based approaches leverage different metrics or modeling assumptions. K-Nearest Neighbors (KNN) based methods, for example, directly quantify OODness by measuring the distance of a test sample to its k-nearest neighbors within the ID training data's feature space, offering a non-parametric alternative to Gaussian models. In a more modern context, (112) introduced PixOOD for pixel-level OOD detection, which, while operating at a finer granularity, fundamentally relies on distances. It extracts pixel/patch feature representations and builds a 2D projection space where distances to multiple class etalons (learned prototypes) are used to model complex intra-class variability and identify OOD pixels. This demonstrates how distance-based principles can be adapted for fine-grained OOD detection by modeling ID distributions with multiple prototypes rather than a single centroid.

In summary, feature-space distance-based methods represent a crucial evolution in OOD detection, moving beyond simple output probabilities to leverage the richer in-

formation in internal representations. They have progressed from initial explorations of Mahalanobis distance on existing features to sophisticated techniques that actively regularize, learn, or project OOD-discriminative feature spaces. While these methods have shown promise in quantifying an input's deviation from the in-distribution manifold, challenges persist, including the sensitivity to the quality of learned features, the computational cost associated with training-time regularization, and the inherent difficulties of distance metrics in high-dimensional spaces. Future directions include developing more flexible models for feature distributions (e.g., non-parametric density estimation or advanced mixture models), integrating self-supervised learning to learn more robust features, and exploring adaptive distance metrics that can better capture complex manifold structures. Furthermore, the utility of these distance-based OOD scores is increasingly recognized in broader uncertainty quantification frameworks, such as their application as non-conformity scores within Conformal Prediction to provide statistically rigorous guarantees on OOD detection performance (149).

3 Generative and Reconstruction-Based Approaches

3.1 Likelihood-Based Deep Generative Models

The detection of Out-of-Distribution (OOD) samples is a critical challenge for deploying reliable deep learning systems. A theoretically principled approach to OOD detection involves leveraging deep generative models to learn the underlying data distribution of in-distribution (ID) samples. The fundamental premise is that samples originating from outside this learned distribution, i.e., OOD samples, should exhibit a significantly lower likelihood or probability density under the model trained exclusively on ID data. This allows for their identification based on their deviation from the learned ID manifold, offering a theoretically grounded method for uncertainty quantification. Early methods primarily explored Variational Autoencoders (VAEs) (**?**) and Generative Adversarial Networks (GANs) (**?**) for this purpose. VAEs provide an estimate of the data log-likelihood (or a lower bound, ELBO), while GANs can be adapted to provide density

estimates or use discriminator scores as proxies for typicality.

Initially, the intuitive strategy was to directly use the raw likelihood or a related reconstruction error from a trained generative model as an OOD score. However, this straightforward application often proved problematic. Raw likelihood scores from deep generative models can be inherently misleading, frequently assigning higher likelihoods to certain OOD samples than to some ID samples (**?**). This counter-intuitive phenomenon, observed across various model architectures and datasets, stems from several factors. Deep generative models, particularly in high-dimensional spaces, may learn spurious low-level correlations or assign high probabilities to simple, out-of-distribution inputs that lie in regions of the input space not well-constrained by the ID data. This can be exacerbated by the "Gaussian Annulus Theorem," where in high dimensions, the typical set of data (where most of the probability mass lies) may not intersect with the region of highest density, leading to OOD samples with high likelihood but low typicality (12). For VAEs, the Evidence Lower Bound (ELBO) is only a lower bound to the true log-likelihood, and a tighter bound does not necessarily correlate with better OOD detection performance. Even models capable of exact likelihood estimation, such as Normalizing Flows (NFs) (**?**) and Autoregressive models (e.g., PixelCNNs), suffer from this issue, often assigning higher likelihoods to less complex OOD images than to complex ID images, further demonstrating that raw likelihood alone is an unreliable indicator of semantic OODness (164).

To address the unreliability of raw likelihood scores, more robust statistical measures were introduced, marking an evolution from simple density estimation to more refined statistical tests. A pivotal advancement was proposed by (**?**), who demonstrated that comparing the likelihood of a sample under the learned ID distribution to its likelihood under a simpler, "null" model provides a significantly more effective OOD score. Their work introduced the concept of likelihood ratio tests for OOD detection, where the ratio of the likelihood under a complex ID generative model (e.g., VAE or GAN) to that under a baseline model (e.g., a simple Gaussian distribution or a model trained on diverse OOD data) serves as a robust discriminator. This approach effectively normalizes the raw likelihood, focusing on how *much better* the ID model explains the data compared to a

general or OOD model, thereby improving OOD discrimination and mitigating the issues associated with misleading raw likelihood values.

Beyond simple likelihood ratios, other sophisticated statistical approaches have emerged to leverage generative models more effectively. (12) introduced Density of States Estimation (DoSE), an unsupervised method that moves beyond direct model probabilities. Inspired by statistical physics, DoSE evaluates the "typicality" of an input by analyzing multiple summary statistics (e.g., negative log-likelihood, L2 norm of latent features) derived from a pre-trained generative model. Instead of directly comparing likelihoods, DoSE trains non-parametric density estimators (like Kernel Density Estimation or one-class SVMs) on the distribution of these statistics for ID data. An OOD sample is then identified if its derived statistics are atypical under these learned distributions, providing a more robust measure of OODness that accounts for the high-dimensional nature of the problem and the shortcomings of raw likelihood.

Furthermore, Normalizing Flows, which provide exact and tractable likelihoods, have seen increasing application in OOD detection, often by modeling densities in feature spaces rather than raw pixel space. For instance, (4) proposed Deep Residual Flow, a novel flow architecture that learns the residual distribution from a base Gaussian distribution, improving OOD detection by modeling feature activations. Similarly, (157) investigated feature density estimation via Normalizing Flows as a fully unsupervised, post-hoc method. By training a lightweight NF model on the feature representations of a pre-trained classifier, they demonstrated strong results for far-OOD detection, highlighting that while raw input likelihoods can be problematic, density estimation in a more semantically meaningful feature space can be highly effective.

The evolution from simple density estimation to more refined statistical measures like likelihood ratio tests and typicality-based approaches, coupled with the broader application of diverse generative architectures like Normalizing Flows, marks a significant progression in the field. While likelihood-based methods offer a theoretically grounded approach to OOD detection, challenges persist in accurately modeling complex, high-dimensional data distributions and ensuring that likelihood estimates genuinely correlate

with semantic OODness across diverse scenarios. Future work continues to explore more robust generative architectures, improved background models for ratio tests, and sophisticated statistical tests to further bridge the gap between theoretical soundness and practical efficacy in real-world OOD detection tasks.

3.2 Reconstruction Autoencoders: Advancements and Limitations

Traditional reconstruction autoencoders (AEs) were initially considered a promising avenue for Out-of-Distribution (OOD) detection, operating under the intuitive assumption that models trained exclusively on in-distribution (ID) data would struggle to reconstruct novel OOD inputs, leading to higher reconstruction errors. However, this foundational premise frequently faltered due to a fundamental design paradox: the inherent capacity of deep autoencoders to generalize and effectively reconstruct diverse novel inputs (**?**). This powerful generalization, while beneficial for tasks like denoising or data compression, often meant that OOD samples yielded reconstruction errors comparable to, or even lower than, ID samples (12). Consequently, the simple reconstruction error proved to be an unreliable metric for distinguishing ID from OOD data, severely limiting the practical utility of early AE-based methods in safety-critical applications. This critical limitation necessitated a significant re-evaluation and sophisticated re-engineering of their underlying principles to transform them into reliable OOD detectors.

Early attempts to leverage reconstruction error for OOD detection, including simpler methods like Principal Component Analysis (PCA) for dimensionality reduction and subsequent reconstruction, often faced this challenge (68). The core problem was designing autoencoders that could learn a sufficiently tight manifold of ID data without inadvertently developing the capacity to reconstruct novel patterns. This challenge spurred significant advancements that fundamentally rethought the autoencoder's objective and architecture, moving beyond raw pixel reconstruction and simple L2 error.

One pivotal advancement has been the shift from pixel-level reconstruction to *semantic feature reconstruction*. Reconstructing raw pixels demands high expressiveness from

the autoencoder, which can inadvertently generalize to OOD inputs. Instead, modern approaches often focus on reconstructing robust, high-level features extracted from pre-trained models. This strategy aligns with broader trends in OOD detection that leverage discriminative feature spaces (**?**). For instance, (2) exemplifies this by reconstructing Activation Vectors (AVs) from the penultimate layer of a pre-trained classifier. This strategic shift simplifies the autoencoder's task to lower-dimensional, semantically relevant features, ensuring that the autoencoder's learning is concentrated on the abstract, task-specific characteristics of ID data. Deviations in reconstructing these semantic features are thus more indicative of OODness than pixel-level discrepancies, which might be influenced by superficial similarities.

Concurrently, to prevent the latent space from accommodating novel patterns, methods have increasingly enforced a *maximally compressed or regularized latent space* for ID samples. This is achieved through regularization losses during training that actively restrict ID latent features to a compact, known domain. Variational Autoencoders (VAEs), a class of generative models that learn a latent distribution, inherently aim for a more structured latent space, which can be leveraged for OOD detection. For example, (20) integrates VAEs into an Inductive Conformal Anomaly Detection (ICAD) framework for real-time OOD detection in cyber-physical systems. Here, the VAE's ability to reconstruct ID data from its learned latent space, combined with Deep Support Vector Data Description (SVDD) for learning a minimum-volume hypersphere, effectively enforces a tight ID manifold. Similarly, (**?**) explores the use of self-attention within VAEs to learn more discriminative and compact latent representations specifically for anomaly detection. (2) also enforces this explicit constraint, ensuring that any input whose latent representation falls outside this tightly defined space is likely OOD. This contrasts sharply with earlier autoencoders that allowed latent spaces to form organically, often encompassing regions where OOD samples could reside without significant reconstruction penalty.

Furthermore, to address the challenge of recovering significant information from an extremely compressed latent space in a single step, (2) proposes a novel *layerwise decomposition for incremental information recovery*. This "data certainty decomposition"

framework factorizes the probability of an input being ID into a product of conditional probabilities, employing a series of decoders. Each decoder is specifically designed to recover information lost after *each individual encoding layer*, rather than a single decoder attempting to recover all accumulated loss from the final, most compressed latent representation. This incremental recovery mechanism enhances the autoencoder's ability to faithfully reconstruct ID samples while remaining highly sensitive to OOD deviations at various levels of abstraction.

Finally, to overcome the issue of standard L2 reconstruction error being an unreliable uncertainty measure (often yielding misleadingly small errors for OOD samples due to smaller activation magnitudes), (2) introduces the *Normalized L2 Distance (NL2)*. This novel metric normalizes the reconstruction by the input's norm, effectively eliminating the confounding influence of feature magnitude and providing a more robust and reliable measure of reconstruction accuracy. The need for more robust scoring functions for reconstruction error is also echoed in works like (68), which demonstrates that even a simple regularized PCA-based reconstruction error can significantly improve OOD detection when fused with other scoring functions, highlighting that raw reconstruction error often requires refinement or combination to be effective. These collective innovations directly address the traditional flaws of reconstruction autoencoders, transforming them into more reliable OOD detectors by ensuring that reconstruction error truly reflects OODness rather than merely the model's generalization capacity.

While these advancements, exemplified by works like (2) and (20), significantly revitalize reconstruction autoencoders for OOD detection, inherent challenges persist. The reliance on a pre-trained classifier for extracting Activation Vectors, as in (2), means the method's effectiveness is intrinsically tied to the quality, robustness, and potential biases of that classifier. If the feature extractor itself is not robust to certain distribution shifts, the OOD detector built upon it will inherit these limitations. Furthermore, defining what constitutes a "maximally compressed" latent space and ensuring its boundaries are sufficiently robust to all possible ID variations, while still being tight enough to reject all OOD, remains an intricate balance. Overly restrictive latent spaces might misclassify

complex ID samples as OOD, while overly permissive ones risk the original generalization problem. The computational overhead introduced by multi-decoder architectures and complex regularization also needs consideration for real-time applications, particularly in resource-constrained environments like those discussed in (20). Future research could explore adaptive mechanisms for latent space compression, investigate more sophisticated theoretical frameworks for quantifying the "OODness" reflected by reconstruction errors in highly complex, high-dimensional data, and develop more robust and adaptive thresholding strategies for reconstruction-based scores, potentially incorporating fusion with other OOD signals as suggested by (68).

3.3 Energy-Based Models for OOD Detection

Energy-Based Models (EBMs) have emerged as a theoretically principled and increasingly effective framework for Out-of-Distribution (OOD) detection, offering a direct approach to modeling the underlying data distribution. At their core, EBMs define a probability distribution over inputs x using an energy function $E(x)$ as $p(x) = \frac{\exp(-E(x))}{Z}$, where $Z = \int \exp(-E(x))dx$ is the intractable partition function (**?**). In this paradigm, in-distribution (ID) samples are characterized by low energy values, indicating high likelihood under the learned distribution, while OOD samples are assigned high energy values, signifying low likelihood. This direct modeling of data likelihood provides a more interpretable and robust measure of OODness compared to relying on proxy metrics like maximum softmax probabilities, which often exhibit overconfidence on OOD inputs (**?**).

A foundational insight into the connection between classifiers and EBMs was provided by Grathwohl_etal_2019, who demonstrated that a standard classifier's output logits can be interpreted as an unnormalized negative energy function. This perspective paved the way for explicitly leveraging energy functions for OOD detection. Building on this, Liu_etal_2020 pioneered the use of Energy-based Models for OOD detection by defining the energy function directly from the output logits of a standard neural network classifier. Their significant contribution lay in developing specialized training objectives to explicitly learn these energy landscapes. This typically involves a contrastive learning

approach, where the model is trained to push the energy of ID samples to be low while simultaneously increasing the energy of OOD samples. The challenge of the intractable partition function Z is often circumvented during training by employing techniques like Stochastic Gradient Langevin Dynamics (SGLD) to sample from the model's distribution and approximate gradients, effectively optimizing the unnormalized density ratio rather than the absolute density (**?** 80). This direct optimization for OOD discrimination offers a more theoretically grounded and robust measure than post-hoc methods.

While many EBM approaches implicitly handle the intractable partition function through contrastive learning and sampling, peng20243ji directly addresses this challenge by proposing ConjNorm, a method for tractable density estimation for post-hoc OOD detection. Their work introduces a novel theoretical framework grounded in Bregman divergence, extending density considerations to the exponential family of distributions. Crucially, ConjNorm devises an unbiased and analytically tractable estimator for the partition function using a Monte Carlo-based importance sampling technique, providing a principled way to estimate true data density without strong distributional assumptions. This represents a significant advancement by offering a direct solution to a core theoretical hurdle in EBMs.

Subsequent research has explored diverse strategies to enhance EBMs for OOD detection. lafon2023w37 introduced HEAT (Hybrid Energy Based Model in the Feature Space), a novel post-hoc method that refines existing OOD detectors (e.g., GMMs, energy logits) by complementing them with a data-driven residual EBM. HEAT uses the EBM framework to compose several energy terms from different refined priors, allowing for accurate and robust ID density estimation without requiring external OOD samples for training. This demonstrates how EBMs can be integrated to correct biases and enhance the expressiveness of other OOD scoring functions.

Furthermore, EBMs have been adapted to address specific challenges in OOD detection. choi202367m proposed a "balanced energy regularization loss" to tackle the problem of imbalanced auxiliary OOD data, which is often overlooked in methods like Outlier Exposure. Their approach adaptively applies larger regularization to auxiliary samples from

majority classes, ensuring a more effective energy landscape shaping. Similarly, in the context of Class-Incremental Learning (CIL), miao20246mk introduced Bi-directional Energy Regularization (BER). BER mitigates biases in CIL models by using energy loss functions to enlarge decision boundaries for new classes (pushing OOD away) and boost confidence for old classes (preventing old ID from being misclassified as OOD), showcasing EBMs' utility in dynamic learning environments.

Beyond direct energy minimization, energy functions can also guide adaptive training. hofmann2024gnx introduced Hopfield Boosting, an OOD detection framework that leverages Modern Hopfield Energy (MHE) to adaptively sample "weak learners" from auxiliary outlier datasets that are hard to distinguish from ID data. By incorporating an MHE-based energy function into the training loss, this method explicitly sharpens the decision boundary between ID and OOD data, demonstrating a sophisticated use of energy to improve outlier exposure strategies.

The scalability of EBMs to modern deep learning architectures has also been a focus. Ming_etal_2023 extended the EBM framework by demonstrating how to effectively fine-tune large pre-trained models, such as vision transformers, for energy-based OOD detection. This approach leverages the rich, generalizable representations learned by these powerful models, addressing the challenge of robust OOD detection in complex, high-dimensional data settings and highlighting the adaptability of the EBM paradigm.

In summary, EBMs provide a theoretically sound framework for OOD detection by directly modeling data likelihood through an energy function. The evolution of EBMs for OOD detection has progressed from foundational insights into their connection with classifiers (**?**) and pioneering contrastive training strategies (**?**), to more sophisticated approaches that directly address the partition function intractability (104), integrate with and refine other OOD methods (80), adapt to specific learning challenges like imbalanced OOD data or incremental learning (39; 158), and leverage energy functions for adaptive training (116). Their ability to integrate with large pre-trained models further underscores their potential for robust OOD detection in real-world applications (**?**).

Future research in EBMs for OOD detection could explore more advanced tech-

niques for approximating or tractably estimating the partition function in complex, high-dimensional settings, potentially drawing from advancements in score-matching or normalizing flows. Investigating adaptive energy functions that can dynamically adjust to different types of OOD shifts or domain contexts, perhaps through hypernetworks, could also yield more versatile detectors. Furthermore, exploring the interplay between EBMs and generative modeling to synthesize highly informative OOD examples for contrastive training, or to learn more expressive energy landscapes, remains a promising avenue.

4 Training-Time Strategies and Robust Representation Learning

4.1 Outlier Exposure and Virtual Outlier Synthesis

The challenge of deploying deep learning models in open-world environments necessitates robust mechanisms for identifying Out-of-Distribution (OOD) inputs. A significant advancement in this area is the Outlier Exposure (OE) paradigm, where auxiliary OOD data is incorporated during model training to explicitly teach the model to distinguish novel inputs. This approach often frames OOD detection as a binary classification task, differentiating between in-distribution (ID) and OOD data.

The theoretical underpinnings of OE suggest that many methods leveraging OOD training data are asymptotically equivalent to a binary discriminator, with practical differences often stemming from estimation procedures and the specific choice of auxiliary OOD data (42). Early implementations of OE demonstrated its effectiveness, but subsequent research has focused on refining its application and addressing inherent limitations. For instance, (39) identified that auxiliary OOD data often exhibits class imbalance, proposing a balanced energy regularization loss to apply stronger regularization to majority OOD classes, thereby enhancing detection performance in diverse tasks like semantic segmentation and long-tailed classification. Addressing the vulnerability of OE to adversarial attacks, (11) introduced Adversarial Learning with inlier and Outlier Exposure

(ALOE), which robustifies detectors by training against both adversarial in-distribution and OOD examples, significantly improving robustness against perturbations.

As OE matured, its application extended to more complex scenarios. In long-tailed recognition, where distinguishing OOD from tail classes is particularly challenging, (35) proposed Calibrated Outlier Class Learning (COCL). This method uses debiased large margin learning and outlier-class-aware logit calibration to explicitly separate OOD samples from both head and tail ID classes, outperforming traditional OE by mitigating class-specific biases. Similarly, (46) introduced EAT, which employs dynamic virtual labels for OOD data and context-rich tail class augmentation to improve OOD detection in long-tailed settings, demonstrating that strong inlier classification does not automatically imply good OOD detection.

Despite its successes, a critical limitation of OE is the reliance on the availability and diversity of *real* auxiliary OOD data. Collecting sufficiently diverse and representative OOD datasets is often impractical or impossible, leading to a shift towards Virtual Outlier Synthesis (VOS). VOS addresses this data scarcity by generating synthetic outliers, thereby overcoming the dependence on real OOD datasets and enhancing model robustness.

Early forays into VOS, such as Mixture Outlier Exposure (MixOE) by (18), generated virtual outliers by mixing ID and auxiliary data. This approach was particularly effective for fine-grained OOD detection, where novel inputs are semantically similar to ID data and require a broader coverage of the feature space. Building on the need for diversity, (72) introduced Diverse Outlier Sampling (DOS), a strategy that selects diverse and informative outliers from auxiliary datasets by combining clustering on normalized features with uncertainty-based selection. This aimed to shape a globally compact decision boundary, improving upon biased greedy sampling. Further advancing this, (85) proposed `diverseMix`, a diversity-induced mixup strategy with theoretical guarantees, which generates semantically distinct synthetic outliers through dynamic interpolation, provably enhancing the diversity of the auxiliary set.

More sophisticated VOS methods have emerged, generating synthetic outliers directly

from in-distribution data or leveraging advanced generative models. (90) introduced Virtual Outlier Smoothing (VOSo), which constructs auxiliary OOD samples by perturbing semantic regions of ID samples in the *image space*, using Class Activation Maps (CAMs). Crucially, VOSo assigns dynamic soft labels based on the perturbation extent, creating a smoother decision boundary and more nuanced uncertainty estimation than traditional uniform OOD labels. Similarly, (66) (MixOOD) also utilized Mixup-based strategies to generate augmented images as auxiliary OOD data, demonstrating improved distinction between ID and OOD samples. (99) explored a "negative branch" method with directional regularization and OOD training data, which implicitly functions as a form of virtual outlier generation to enhance anomaly detection.

The advent of large pre-trained models, particularly Vision-Language Models (VLMs), has opened new avenues for VOS. (82) proposed Outlier Label Exposure (OLE) for zero-shot OOD detection, which generates textual outlier prototypes by clustering and refining auxiliary outlier class labels. This effectively synthesizes OOD knowledge in the language domain, enhancing VLM safety without extensive training. Complementing this, (96) introduced `NegPrompt`, a method that learns transferable negative prompts for each ID class using only ID data. These negative prompts implicitly define OOD boundaries, enabling open-vocabulary OOD detection by leveraging the VLM's semantic understanding. (41) (GL-MCM) further explored VLM capabilities by combining global and local concept matching for zero-shot OOD, implicitly handling multi-object OOD scenarios by leveraging local features to overcome contamination of global features. (109) developed Self-Calibrated Tuning (SCT) for VLMs, which adaptively balances ID classification and OOD regularization by leveraging ID-irrelevant local context as surrogate OOD data, addressing the issue of spurious OOD features.

Generative models, especially diffusion models, have also been harnessed for VOS. (44) introduced DiffGuard, a semantic mismatch-guided OOD detection method that uses pre-trained diffusion models. It synthesizes new images conditioned on an input and its predicted label, identifying OOD samples by measuring the dissimilarity between the original and synthesized images. This approach leverages the conditional generation

capabilities of diffusion models to highlight semantic contradictions, overcoming scalability issues of prior generative methods.

The VOS paradigm has also extended to specialized domains and multimodal inputs. For pixel-wise OOD detection in semantic segmentation, (28) (ObsNet+LAA) generates OOD-like training data via local adversarial attacks, simulating unknown objects to train an auxiliary observer network. Similarly, (15) (RPL) utilizes Outlier Exposure with synthetic OOD data to learn residual anomaly patterns without retraining the base segmentation model. In 3D LiDAR-based object detection, (81) generates synthetic OOD objects by perturbing known ID object categories, addressing data scarcity in this domain. For multimodal OOD detection, (100) introduced Nearest Neighbor Prototype-based Mixup (NP-Mix) as part of their Agree-to-Disagree (A2D) algorithm, generating outliers by leveraging nearest neighbor class prototypes to explore broader feature spaces. Building on this, (113) proposed Dynamic Prototype Updating (DPU), which dynamically adjusts multimodal prediction discrepancy intensification based on a sample's similarity to its class prototype, accounting for intra-class variability in multimodal data. Finally, (116) introduced Hopfield Boosting, an OE approach that adaptively samples "hard" outliers using a novel energy function derived from Modern Hopfield Networks, further refining the selection of informative synthetic or real outliers.

In conclusion, Outlier Exposure has evolved from a foundational paradigm to a sophisticated framework that explicitly trains models to recognize novel inputs. The critical challenge of OOD data scarcity has driven the field towards Virtual Outlier Synthesis, which leverages advanced techniques like semantic-level interpolation, adversarial generation, and prompt-based synthesis to create diverse and effective training examples. While VOS has significantly reduced the reliance on real OOD datasets and enhanced model robustness across various modalities and tasks, ongoing challenges include ensuring the representativeness of synthetic outliers for truly unknown OOD distributions, scaling complex generation methods, and developing stronger theoretical guarantees for their generalization capabilities.

4.2 Learning Robust and Separable Feature Representations

The intrinsic quality and structured organization of a model's internal feature representations are paramount for effective Out-of-Distribution (OOD) detection. This subsection delves into advanced methodologies that actively engineer the deep neural network's embedding space, aiming to enhance the discriminability between in-distribution (ID) and OOD samples. These approaches collectively improve the inherent OOD robustness of the learned representations by designing a more structured and discriminative embedding space, often by enforcing explicit geometric separation, creating more compact and well-defined ID clusters, or refining feature transformations.

A significant line of research leverages the intrinsic properties of deep neural networks, particularly the phenomenon of Neural Collapse (NC), to enforce explicit geometric separation between ID classes and push OOD samples away. Neural Collapse describes the convergence of features within each class to their respective class means, and these class means to a simplex equiangular tight frame (ETF) structure in the terminal phase of training. Building on this, (27) introduced NECO, a post-hoc method that capitalizes on a newly observed property dubbed ID/OOD Orthogonality (NC5). This property posits that OOD data tends to become increasingly orthogonal to the principal component space spanned by ID class means. NECO exploits this by projecting features onto this ID-derived principal component space, using the projection magnitude as an OOD score. A smaller projection magnitude indicates higher OOD likelihood. Extending this concept, (84) proposes a novel separation loss (`LSep`) that actively constrains OOD features to reside in a subspace orthogonal to the principal subspace of ID features, which is implicitly defined by the final layer's weights. This approach moves beyond merely observing orthogonality to explicitly enforcing it during training, typically by leveraging auxiliary OOD data to push their features into distinct, non-overlapping dimensions. An earlier, more general effort towards compact ID representations, (7) proposed embedding ID samples into a low-dimensional space forming a union of 1-dimensional subspaces, arguing that such a highly constrained representation inherently makes it less likely for OOD samples to occupy ID regions. While Neural Collapse-based methods offer strong

theoretical underpinnings for explicit separation, their robustness to subtle, near-OOD shifts that might not perfectly align with orthogonal subspaces remains a critical area for further investigation, as these shifts might still project significantly onto the ID subspace.

Beyond direct geometric enforcement, feature transformation and subspace learning techniques are employed to enhance separability, often addressing the "curse of dimensionality" that can plague distance-based methods in high-dimensional feature spaces. Traditional linear dimensionality reduction methods often struggle to capture the complex non-linear relationships inherent in deep features. (5) introduced RankFeat, a post-hoc method that uses Singular Value Decomposition (SVD) to identify and remove a dominant rank-1 component from high-level features. This spectral manipulation implicitly refines the feature space by mitigating the over-confidence induced by this component in OOD samples, effectively "flattening" the feature manifold. Addressing the limitations of purely linear transformations, (111) proposes Kernel PCA (KPCA) for OOD detection, devising novel non-linear mappings like Cosine Mapping (CoP) and Cosine-Gaussian Mapping (CoRP). These mappings explicitly transform features into a space where ID and OOD data become linearly separable, overcoming the ineffectiveness of conventional PCA on raw features, a challenge also noted by (68) which explored regularized PCA reconstruction errors. To directly combat the curse of dimensionality, (53) proposes Subspace Nearest Neighbor (SNN), which regularizes the model and its feature representation by leveraging the most relevant subset of dimensions (i.e., subspace) during training. This subspace learning yields more distinguishable distance measures between ID and OOD data. Similarly, (162) introduces a data structure-aware approach using a novel "tangent distance" that maps high-dimensional features to the manifold of ID samples. By directly computing the Euclidean distance between samples and the nearest submanifold space (linear approximation of local regions on the manifold), it mitigates the sensitivity of distances to high dimensionality, proposing that OOD samples are relatively far from the ID manifold. The computational overhead and the challenge of selecting optimal non-linear kernels or relevant subspaces for diverse OOD scenarios represent practical considerations for these transformation-based methods.

Another significant direction involves refining ID clusters through prototype-based learning and mixture models, which aim to capture the nuanced structure of ID data more faithfully. Traditional distance-based OOD methods often oversimplify ID classes by modeling them with a single centroid, failing to capture intra-class diversity and leading to suboptimal OOD boundaries. (93) addresses this with Prototypic Learning with a Mixture of prototypes (PALM), which models each ID class with multiple prototypes using a mixture of von Mises-Fisher (vMF) distributions in a hyperspherical embedding space. This approach creates more faithful and compact ID clusters, allowing for a more precise definition of ID boundaries by optimizing both a Maximum Likelihood Estimation (MLE) loss and a novel prototype contrastive loss. Extending this concept to multimodal settings, (113) introduces Dynamic Prototype Updating (DPU), which dynamically adjusts multimodal prediction discrepancy intensification based on a sample's similarity to its class prototype. DPU employs Cohesive-Separate Contrastive Training (CSCT) to build a robust representation space and Pro-ratio Discrepancy Intensification (PDI) to balance intra-class cohesion with inter-class separation, thereby enhancing OOD detection in complex multimodal data. Complementing these empirical approaches, (121) provides theoretical insights into how in-distribution labels help OOD detection, particularly for "near OOD" scenarios. Through a graph-theoretic framework and spectral decomposition, they demonstrate that ID labels, by defining supervised connectivity, enable the learning of more discriminative ID representations that facilitate OOD-ID separation, especially when ID data is sparsely connected without labels. While prototype-based methods offer improved fidelity, the challenge of determining the optimal number of prototypes and their sensitivity to noisy ID data remains, and OOD samples falling between distinct ID prototypes can still pose detection difficulties.

Despite these significant advancements in actively structuring and refining feature spaces, a persistent challenge remains in ensuring that these learned representations generalize effectively to truly novel and diverse OOD types. While methods leveraging Neural Collapse offer promising theoretical underpinnings for explicit separation, their robustness to subtle, near-OOD shifts that might not perfectly align with orthogonal subspaces

requires further investigation. Similarly, prototype-based methods, while improving intra-class modeling, still face the inherent difficulty of defining boundaries for the unknown, especially when OOD data falls within the convex hull of ID prototypes. Future research could focus on adaptive feature space shaping techniques that dynamically adjust to the evolving nature of OOD data, perhaps through meta-learning or continuous adaptation mechanisms, to achieve more universally robust and discriminative representations that are less sensitive to the specific characteristics of unseen OOD data.

4.3 Gradient-Based and Neuron-Level Analysis

A significant shift in Out-of-Distribution (OOD) detection research involves delving into the fine-grained internal dynamics of neural networks, leveraging gradient information and individual neuron activations to extract more precise and interpretable OOD signals. This introspection moves beyond aggregate model outputs to understand *how* and *why* a model processes OOD inputs differently.

One prominent direction focuses on abnormalities in gradient-based attribution maps, which reveal how input features influence predictions. (77) introduced GAIA (Gradient Abnormality Inspection and Aggregation), a framework that quantifies the "abnormality" in gradient-based attribution results, observing that OOD samples lead to "meaningless attribution results" with abnormal non-zero density in deeper layers. Building on gradient insights, (47) proposed GradOrth, which identifies OOD data by computing the norm of the gradient projection onto a low-rank subspace deemed important for in-distribution (ID) data, indicating a weak correlation with ID patterns. While these methods are post-hoc, analyzing gradients after training, (11) introduced ALOE (Adversarial Learning with inlier and Outlier Exposure), a training-time strategy that uses adversarial examples generated via gradients to robustify OOD detectors against small input perturbations. ALOE's objective is to promote smoother OOD score functions for ID data and clearer separation for OOD data, directly addressing the robustness aspect through gradient-informed regularization during training.

Another crucial area explores the 'coverage' of neuron activation states by in-distribution

data, revealing deviations from learned patterns. (30) proposed Neuron Activation Coverage (NAC), a novel statistical measure that quantifies how well neuron states are "covered" by ID training data, serving as an uncertainty measure for OOD detection and a metric for OOD generalization. This approach provides a deeper, neuron-centric understanding of OOD phenomena. Complementing this, (9) introduced Batch Normalization Assisted Typical Set Estimation (BATS) with a Truncated BN (TrBN) unit, which rectifies extreme feature activations into their "typical set" to boost OOD detection scores, effectively managing neuron states. (62) further generalized this concept with Variational Rectified Activation (VRA), providing a theoretical derivation for an optimal activation function that not only suppresses abnormally high activations (like TrBN) but also low ones, and amplifies intermediate activations, leading to superior OOD separation. These rectification methods demonstrate a progression from heuristic to theoretically grounded manipulation of neuron activations.

Beyond individual neuron states, statistical analyses of activation patterns also prove effective. (19) developed Neural Mean Discrepancy (NMD), a metric that quantifies the difference between the average activations (neural means) of input examples and the training data across multiple layers. NMD leverages Batch Normalization's running averages for efficiency, providing a lightweight yet powerful OOD signal. More recently, (89) presented SISOM (Simultaneous Informative Sampling and Outlier Mining), a unified approach for active learning and OOD detection that enriches feature representations by weighting neurons based on their gradient contribution to the KL divergence between a uniform distribution and the model's softmax output. This method effectively combines gradient-based saliency with neuron activation analysis to identify unexplored regions and decision boundaries, showcasing a sophisticated integration of these fine-grained internal dynamics.

The collective efforts in this subsection highlight a growing trend towards deeper introspection into model internals. By analyzing gradient-based attribution maps, the coverage of neuron activation states, and employing gradient regularization during training, researchers are developing more precise, interpretable, and robust OOD detection methods.

However, challenges remain in establishing universal patterns for gradient abnormalities or neuron coverage across diverse architectures and OOD types, and in balancing the computational cost of such fine-grained analysis with real-time deployment needs.

5 Advanced OOD Paradigms and Contexts

5.1 Multimodal and Graph-Structured OOD Detection

The landscape of Out-of-Distribution (OOD) detection is rapidly expanding beyond traditional unimodal image data to encompass the complexity of real-world multimodal and graph-structured information. This crucial extension addresses the inherent multimodal nature of many applications and the unique challenges posed by non-Euclidean data.

For graph-structured data, OOD detection presents distinct challenges due to its non-Euclidean nature and the high cost of labeling. Pioneering efforts have focused on unsupervised graph-level OOD. (6) introduced GOOD-D, a novel framework for unsupervised graph-level OOD detection that learns robust in-distribution (ID) patterns through perturbation-free data augmentation and hierarchical graph contrastive learning across node, graph, and group levels. This approach was critical in formalizing the problem and providing a multi-granularity understanding of graph ID data. Building on the need for OOD exposure in graphs, (48) proposed GOLD, which addresses the scarcity of auxiliary OOD data by implicitly generating pseudo-OOD samples through an adversarial latent generation framework, achieving superior performance without real OOD samples. Addressing practical deployment constraints, (105) introduced GOODAT, a test-time graph OOD detection method that operates without access to training data or requiring GNN architecture modifications, leveraging a graph masker and the Graph Information Bottleneck (GIB) principle for unsupervised OOD identification. To provide a unified evaluation framework for this nascent field, (106) presented UB-GOLD, a comprehensive benchmark that unifies unsupervised graph-level anomaly detection and OOD detection across 35 datasets and four distinct scenarios, enabling systematic comparison and analysis of diverse graph OOD methods.

The detection of OOD samples in multimodal settings is gaining traction as real-world data often comprises complementary information from diverse sources like vision, audio, and text. (100) made a significant contribution by introducing MultiOOD, the first dedicated benchmark for multimodal OOD detection, alongside the Agree-to-Disagree (A2D) algorithm. A2D leverages the "modality prediction discrepancy" phenomenon, where softmax predictions across modalities show negligible differences for ID data but significant variability for OOD data, to enhance detection. Extending this concept, (113) proposed Dynamic Prototype Updating (DPU), a plug-and-play framework that addresses the limitation of uniform discrepancy amplification by dynamically adjusting intensification based on a sample's similarity to its class prototype, thereby balancing intra-class cohesion with inter-class separation.

Vision-Language Models (VLMs) have emerged as a powerful paradigm for multimodal OOD detection, particularly in zero-shot and open-vocabulary settings. (41) introduced GL-MCM, which enhances zero-shot OOD detection by combining CLIP's global and local features, offering flexibility in defining ID images in complex, multi-object scenes. Further refining VLM-based approaches, (96) developed NegPrompt, a method that learns transferable "negative prompts" using only ID training data to delineate OOD boundaries, enabling open-vocabulary OOD detection without explicit OOD examples. Addressing the challenge of spurious OOD features that can arise from imperfect foreground-background decomposition in VLMs, (109) proposed Self-Calibrated Tuning (SCT), an adaptive framework that dynamically balances ID classification and OOD regularization based on prediction uncertainty. Beyond VLMs, Large Language Models (LLMs) are being explored for their world knowledge. (60) leveraged LLMs for multi-modal OOD detection by generating descriptive features for ID classes, crucially developing a consistency-based uncertainty calibration method to mitigate LLM hallucinations and prevent performance degradation. This integration of LLMs with VLMs for OOD detection is further contextualized by (101), which provides a comprehensive survey of OOD detection in the VLM/LVLM era, highlighting the integration of related fields and identifying the most demanding challenges.

The expansion of OOD detection to multimodal and graph-structured data marks a significant step towards more holistic and context-rich detection capabilities. While promising advancements have been made in developing new benchmarks and algorithms that exploit inter-modal discrepancies and address the unique challenges of non-Euclidean data, several unresolved issues remain. These include the scalability of multimodal OOD methods to a wider array of modalities beyond vision-language, the robustness of graph OOD detectors to diverse and subtle structural shifts, and the development of theoretically grounded adaptive algorithms that can seamlessly handle the inherent noise and heterogeneity in real-world multimodal and graph data.

5.2 OOD in Specialized Learning Settings

Out-of-distribution (OOD) detection becomes particularly challenging in specialized learning paradigms where inherent data characteristics complicate the distinction between in-distribution (ID) and novel samples. This subsection delves into OOD detection within long-tailed recognition and class-incremental learning, highlighting the unique complexities and tailored solutions developed to ensure OOD robustness in these realistic scenarios.

In **long-tailed recognition (LTR)**, the severe class imbalance creates a pervasive confusion between tail-class ID samples and true OODs. Models often exhibit overconfidence on dominant head classes, leading to OOD samples being misclassified into these categories, while simultaneously treating sparse tail-class instances as anomalies (35; 46; 169). Addressing this requires strategies that either modify the learning objective, augment data, or engineer the representation space to explicitly disentangle tail-class ID from OOD.

One prominent approach involves modifying the learning objective or expanding the label space. (35) introduced Calibrated Outlier Class Learning (COCL), which extends the label space with an explicit outlier class. COCL employs a debiased large margin learning strategy, incorporating OOD-aware tail class prototype learning to prevent tail samples from being mistaken for OOD, and debiased head class learning to mitigate the dominant influence of head classes on OOD samples. This direct manipulation of the

decision boundary in the logit space offers a targeted solution to the class imbalance problem. Complementing this, (39) recognized that even auxiliary OOD data used in outlier exposure can exhibit class imbalance. They developed a balanced energy regularization loss that adaptively applies stronger regularization to auxiliary samples from majority classes, ensuring a more effective learning of OOD boundaries in long-tailed and semantic segmentation tasks.

Another crucial strategy focuses on data augmentation and dynamic outlier adaptation. (46) proposed EAT, a framework that utilizes dynamic virtual labels for OOD data and context-rich tail class augmentation. By overlaying tail-class images onto OOD backgrounds, EAT encourages the model to focus on discriminative foreground features, improving both tail-class generalization and OOD distinction. This data-centric approach contrasts with COCL's loss modifications by enriching the training data itself. Further advancements in dynamic outlier adaptation include (90)'s Virtual Outlier Smoothing (VOSo), which constructs virtual outliers by perturbing semantic regions of ID samples in the image space and assigns them dynamic soft labels. This creates a smoother decision boundary, preventing tail classes from being abruptly classified as OOD. To enhance the diversity of auxiliary OOD data, (85) theoretically demonstrated that increased diversity improves OOD generalization and proposed `diverseMix`, a semantic-level interpolation strategy with dynamic adjustment. Similarly, (72) introduced Diverse Outlier Sampling (DOS), which selects diverse and informative outliers from auxiliary datasets by clustering normalized latent representations. These dynamic sampling and generation techniques, along with adaptive weighting strategies like Hopfield Boosting (116) that prioritize "hard" outliers, collectively contribute to refining the decision boundary in long-tailed settings, ensuring that tail classes are not erroneously flagged as OOD while true anomalies are detected.

Beyond explicit outlier exposure and loss modifications, engineering the representation space is critical. (169) introduced **Representation Norm Amplification (RNA)**, a novel training method that directly addresses the trade-off between LTR classification accuracy and OOD detection performance. RNA decouples ID classification and OOD

detection by leveraging the norm of the representation vector as a dedicated dimension for OOD scoring. It achieves this by training the classifier to minimize classification loss only for ID samples, while simultaneously regularizing to enlarge the norm of ID representations. Crucially, auxiliary OOD samples are used to regularize Batch Normalization (BN) layers, indirectly reducing OOD representation norms and creating a discernible difference in activation ratios and representation norms. This allows for simultaneous high performance in both tasks, overcoming limitations of previous methods. Similarly, (93) proposed Prototypic ALearning with a Mixture of prototypes (PALM), which models each ID class with multiple prototypes using a mixture of von Mises-Fisher distributions. While not exclusively for long-tailed settings, this approach is highly beneficial for capturing the inherent intra-class diversity within sparse tail classes, leading to more faithful embeddings and improved ID-OOD separability. (37) introduced Multi-scale OOD DEtection (MODE), leveraging both global and local representations with an attention-based local propagation mechanism. This multi-scale approach can help distinguish fine-grained tail-class features from OOD noise, especially when global features are ambiguous due to background clutter.

In **class-incremental learning (CIL)**, where models continuously learn new classes over time, maintaining robust OOD performance is a significant hurdle due to catastrophic forgetting of previously learned classes. The challenge lies in adapting to new ID classes without degrading the OOD detection capability for both old and new data distributions. This area has historically been underexplored, but recent work has begun to provide dedicated solutions and benchmarks.

(158) introduced OpenCIL, the first comprehensive benchmark for OOD detection in CIL, highlighting that CIL models exhibit increasing biases towards OOD samples and newly added classes with more incremental steps, leading to decreased OOD detection performance. OpenCIL proposes two frameworks for integrating OOD detection into CIL: post-hoc methods (applying OOD scores on CIL model features) and fine-tuning-based methods (training an additional OOD classifier while freezing the CIL backbone). To mitigate the identified biases, (158) further proposed Bi-directional Energy Regulariza-

tion (BER). BER addresses two key issues: New Task Energy Regularization (NTER) prevents OOD samples from being over-confidently classified into new classes by synthesizing pseudo-OOD samples and enlarging decision margins. Old Task Energy Regularization (OTER) prevents old ID samples from being misclassified as OOD (due to catastrophic forgetting) by boosting prediction confidence for old classes using augmented memory samples. BER provides a targeted solution to the unique challenges of OOD in CIL by explicitly addressing the dynamic nature of the ID distribution.

Another promising direction is the integration of uncertainty quantification methods. (78) proposed Continual Evidential Deep Learning (CEDL), which integrates Evidential Deep Learning (EDL) into a continual learning framework to simultaneously perform incremental object classification and OOD detection. CEDL combines exemplar rehearsal and knowledge distillation with a novel loss function that includes evidential cross-entropy, KL-divergence regularization for new classes, and knowledge distillation. Their findings indicate that evidential vacuity is a good indicator for OOD detection in CIL, while dissonance struggles to distinguish old ID from OOD. This work offers a principled way to estimate and leverage uncertainty for OOD detection in evolving CIL environments.

Beyond these dedicated CIL-OOD methods, several general OOD concepts offer indirect but promising contributions. Techniques that enhance training stability and prevent overconfidence are crucial in CIL. (63)'s Average of Pruning (AoP), which combines model averaging and network pruning, could help mitigate OOD detection instability and overfitting during continuous learning. Similarly, (86)'s Optimal Parameter and Neuron Pruning (OPNP), a training-free method, could reduce overconfidence in CIL models without requiring additional training data, thus improving OOD discrimination. Leveraging generic pre-trained representations, as explored by GROOD (92), might offer a more stable foundation for OOD detection in CIL, as these representations are less susceptible to task-specific catastrophic forgetting compared to features learned from scratch. The dynamic prototype updating mechanism in (113)'s DPU, though developed for multimodal OOD, conceptually aligns with the need to dynamically refine class representations in CIL to maintain stable boundaries for OOD detection.

In conclusion, specialized learning settings like long-tailed recognition have seen significant progress through tailored solutions that address the nuanced interactions between ID and OOD data, often leveraging dynamic outlier adaptation, sophisticated representation learning, and explicit norm amplification. Crucially, the field of OOD detection in class-incremental learning is rapidly maturing, moving from an underexplored area to one with dedicated benchmarks and methods like OpenCIL and BER, and principled uncertainty-aware approaches like CEDL. Future research needs to further integrate these insights, developing dynamic and adaptive OOD detection frameworks that can explicitly handle evolving ID distributions and catastrophic forgetting, ensuring robust OOD performance in truly open-ended, lifelong learning scenarios.

5.3 Leveraging Pre-trained Foundation Models

The emergence of large pre-trained foundation models, such as Vision-Language Models (VLMs) like CLIP (**?**) and Large Language Models (LLMs), has profoundly transformed the landscape of Out-of-Distribution (OOD) detection. These models, with their rich semantic understanding, vast world knowledge, and open-vocabulary capabilities, enable novel zero-shot and open-set OOD detection paradigms, moving towards more adaptable and versatile OOD systems.

The "Vision Language Model Era" marks a significant paradigm shift in OOD detection, as highlighted by (101). Early integration of VLMs for OOD detection focused on leveraging their inherent zero-shot capabilities. For instance, (41) introduced GL-MCM, extending CLIP's Maximum Concept Matching by utilizing both global and local visual-text alignments. This approach provides flexibility in defining in-distribution (ID) images in multi-object scenes, addressing the limitation of methods that assume single, centered objects. Building upon this, (82) proposed Outlier Label Exposure (OLE), a method that explicitly incorporates OOD knowledge by using a large set of diverse auxiliary outlier class labels as pseudo OOD text prompts for VLMs. OLE learns refined "outlier prototypes" and generates "hard outlier prototypes" to calibrate decision boundaries, overcoming the limitations of purely ID-label-based zero-shot methods that lack explicit OOD knowledge.

Further advancing this direction, (96) introduced NegPrompt, a prompt learning-based approach that learns transferable "negative prompts" for each ID class using *only* ID training data. These negative prompts implicitly define OOD boundaries by representing characteristics contrary to ID classes, enabling open-vocabulary OOD detection without the need for any external outlier data or additional encoders, which is a significant step towards data-efficient and generalizable OOD systems.

Beyond VLMs, Large Language Models (LLMs) are increasingly leveraged for their extensive world knowledge to generate synthetic OOD exposure. (60) explored using LLMs to generate descriptive features for ID classes to enhance multimodal OOD detection. Crucially, they identified and addressed the LLM "hallucination" problem by proposing a novel consistency-based uncertainty calibration method. This method selectively integrates reliable LLM knowledge, preventing performance degradation caused by unfaithful LLM generations. Taking this concept further, (103) introduced Envisioning Outlier Exposure (EOE), which directly uses LLMs to *envision* and generate potential outlier class labels based on visual similarity to ID classes. EOE designs task-specific LLM prompts for far, near, and fine-grained OOD scenarios, effectively creating synthetic OOD labels without access to any real OOD data, thereby providing a powerful form of "outlier exposure" to VLMs.

The integration of foundation models also necessitates refining their behavior for OOD detection. (109) proposed Self-Calibrated Tuning (SCT) for VLMs, a novel framework designed to mitigate the issue of "spurious OOD features" that arise from VLMs' imperfect foreground-background decomposition. SCT adaptively adjusts the balance between ID classification and OOD regularization based on prediction uncertainty, making VLM-based OOD detection more robust to internal model limitations. In a broader context, the field is also expanding to multimodal OOD detection, where foundation models can play a crucial role. (100) introduced the MultiOOD benchmark and the Agree-to-Disagree (A2D) algorithm, which leverages "modality prediction discrepancy" for OOD detection across multiple modalities. While not exclusively VLM/LLM-focused, this work highlights the growing need for robust multimodal OOD, a domain where foundation models

are inherently well-suited to integrate and leverage diverse information streams. Finally, the increasing adoption of foundation models for OOD detection has led to dedicated benchmarks, such as the one presented by (172), which aims to properly assess the performance of these large pre-trained models in realistic yet harder OOD tasks, confirming their benefits and guiding future research into their fine-tuning strategies.

In conclusion, the advent of pre-trained foundation models has opened a new frontier in OOD detection, moving beyond traditional methods that often rely on explicit OOD data or complex architectural modifications. These models' inherent semantic understanding and vast knowledge enable sophisticated zero-shot and open-vocabulary OOD paradigms through techniques like learning transferable negative prompts, leveraging LLMs for envisioned outlier exposure, and self-calibrated tuning of VLMs. However, challenges remain in ensuring the robustness of LLM-generated information, fully integrating multimodal cues, and developing comprehensive benchmarks that capture the full spectrum of OOD scenarios for these powerful, general-purpose models.

6 Real-World Applications and Deployment Challenges

6.1 OOD in Medical Imaging and Healthcare

The deployment of artificial intelligence (AI) in medical imaging and healthcare necessitates robust out-of-distribution (OOD) detection capabilities, as misinterpreting novel or anomalous inputs can have severe, life-threatening consequences for patient well-being. AI-powered clinical decision support systems, used for tasks like disease diagnosis and anomaly detection in medical scans, must reliably identify when an input falls outside their training distribution to prevent erroneous predictions. This domain presents unique challenges, including the high dimensionality of medical images, inherent class imbalance in rare disease detection, and the stringent requirement for robust performance on subtle OOD shifts that might indicate critical pathologies.

To address the foundational understanding of OOD in this high-stakes field, hong2024xls provide a comprehensive survey, establishing a critical taxonomy for distributional shifts

in medical imaging. They delineate seven key factors causing OOD shifts and categorize them into semantic, covariate, and contextual shifts, clarifying the complex landscape and inconsistent terminology that hinders systematic research. Empirically validating the limitations of current approaches, vasiliuk20233w9 expose the severe shortcomings of state-of-the-art OOD detection methods when applied to 3D medical image segmentation. Their work introduces a novel, publicly available benchmark that simulates diverse clinical OOD scenarios and, notably, demonstrates that a simple Intensity Histogram Features (IHF) baseline often outperforms complex deep learning methods, underscoring the profound challenges posed by 3D medical data and the need for more tailored solutions.

Early efforts to adapt general OOD detection techniques to medical imaging revealed significant performance discrepancies. berger20214a3 conducted a comparative study of confidence-based OOD methods on chest X-rays, finding that methods performing well on standard computer vision benchmarks often failed in the medical context. Their analysis highlighted ODIN as a robust method due to its input perturbation mechanism, while Mahalanobis distance, a strong performer in general vision, proved ineffective in medical imaging due to less separable feature spaces. Directly addressing this limitation, anthony2023slf critically re-evaluated the use of Mahalanobis distance for OOD detection in medical imaging. Through a detailed layer-wise analysis, they demonstrated that the optimal detection layer is highly dependent on the specific OOD pattern, challenging previous assumptions. They then proposed Multi-branch Mahalanobis (MBM), a novel framework that significantly enhances OOD detection by employing multiple depth-specific detectors, showcasing a tailored solution that improves reliability for identifying unexpected anomalies like pacemakers or subtle demographic shifts.

Beyond adapting existing discriminative methods, novel generative approaches have emerged to tackle the high dimensionality and complexity of 3D medical data. graham20232re introduced an unsupervised 3D OOD detection method leveraging Latent Diffusion Models (LDMs). This innovative approach overcomes the memory and resolution limitations of prior generative models, enabling the generation of high-resolution spatial anomaly maps. Such capabilities are crucial for tasks like identifying unexpected

tumors, lesions, or other pathologies in volumetric scans, where precise localization is paramount for clinical utility and ensuring the overall reliability of AI-powered diagnostic systems.

Despite these advancements, several challenges persist. The generalizability of OOD methods across the vast spectrum of medical imaging modalities, anatomical regions, and diverse OOD types remains an open problem. There is a continuous need for more comprehensive and clinically relevant benchmarks that capture the subtlety and variability of real-world OOD shifts. Furthermore, integrating these detection mechanisms seamlessly into clinical workflows, coupled with robust explainability and uncertainty quantification, is essential for fostering trust and enabling the safe and effective deployment of AI in patient care.

6.2 OOD for Autonomous Systems and Cyber-Physical Systems

The safe and reliable deployment of autonomous systems, encompassing self-driving vehicles, robotics, and industrial cyber-physical systems (CPS), critically depends on their ability to detect and appropriately react to Out-of-Distribution (OOD) events. In these dynamic, open-world environments, encountering unforeseen objects, sensor malfunctions, or adversarial attacks can lead to catastrophic failures if not promptly identified. This subsection delves into the specialized advancements in OOD detection that address the stringent requirements of real-time performance, robust multimodal sensor fusion, and the nuanced handling of diverse OOD events in such safety-critical applications.

Ensuring the trustworthiness of learning-enabled components in CPS necessitates OOD detection with strong statistical guarantees and real-time capabilities. Early work by (20) addressed this by integrating Variational Autoencoders (VAEs) and Deep Support Vector Data Description (SVDD) within an Inductive Conformal Anomaly Detection (ICAD) framework, providing well-calibrated false alarm rates for high-dimensional sensor inputs. Building on such foundational guarantees, (17) introduced iDECODe, which leverages in-distribution equivariance for conformal OOD detection, offering bounded false detection rates. Extending this to dynamic environments, (87) proposed CODiT for OOD

detection in time-series (dependent) data within CPS, utilizing temporal equivariance and Fisher's method for robust, guaranteed false alarm rates. A crucial practical concern in safety-critical systems is managing false positives; (118) tackled this with a human-in-the-loop framework that adaptively updates OOD detection thresholds using expert feedback and provides theoretical guarantees on false positive rates, even under distribution shifts. Beyond mere statistical detection, (29) argued that traditional OOD detection is insufficient for safety-critical contexts, advocating for Out-of-Model-Scope (OMS) detection, which focuses on identifying inputs that lead to actual model errors, thereby providing a more direct measure of safety. Robustness against malicious inputs is also paramount; (11) introduced ALOE (Adversarial Learning with inlier and Outlier Exposure), a training-time strategy that robustifies OOD detectors against both adversarial in-distribution and OOD examples, a critical defense against cyber-attacks in CPS. For continually evolving autonomous systems, (78) proposed Continual Evidential Deep Learning (CEDL), enabling simultaneous incremental learning of new classes and OOD detection, crucial for systems operating in open-ended environments.

A significant challenge in autonomous systems is the effective integration of heterogeneous sensor data, such as LiDAR, camera, and radar, for robust OOD detection. Traditional unimodal OOD methods often fail to leverage the complementary information across modalities, which is vital for distinguishing subtle OOD events from sensor noise or adverse environmental conditions. Addressing this, (100) introduced MultiOOD, the first dedicated benchmark for multimodal OOD detection, alongside the "Agree-to-Disagree" (A2D) algorithm and "Nearest Neighbor Prototype-based Mixup" (NP-Mix) for outlier synthesis. Their work demonstrated that leveraging modality prediction discrepancies significantly enhances OOD performance, providing a foundational step for multimodal OOD, although primarily evaluated on video-based action recognition. Directly targeting autonomous driving, (180) proposed "Feature Mixing," an extremely simple and fast multimodal outlier synthesis method for OOD detection and segmentation, specifically for image and point cloud data. This method, which randomly swaps feature dimensions between modalities, achieves state-of-the-art performance with significant speedups over

prior methods like NP-Mix, making it highly practical for real-time applications. They also introduced CARLA-OOD, a challenging synthetic multimodal dataset for OOD segmentation in driving scenarios. Further specializing in 3D perception, (81) revisited OOD detection in LiDAR-based 3D object detection, proposing a lightweight post-hoc method that integrates features from the backbone, bounding box parameters, and output logits of a fixed 3D object detector. Crucially, they introduced a novel synthetic OOD generation strategy by perturbing known ID objects and established a new nuScenes OOD benchmark, providing a more realistic evaluation protocol for unknown foreground objects in autonomous driving. For multimodal intent understanding, (88) proposed MIntOOD, integrating weighted feature fusion with multi-granularity representation learning for both classification and OOD detection, highlighting the need for context-aware OOD in complex autonomous tasks.

The advent of large pre-trained foundation models, particularly Vision-Language Models (VLMs), offers new avenues for open-world OOD detection in autonomous systems by leveraging their vast semantic understanding. (171) explored language-enhanced latent representations for OOD detection in autonomous driving, using the cosine similarity of image and text representations encoded by CLIP. This approach improves the transparency and controllability of latent encodings, demonstrating superior performance on realistic driving data compared to traditional vision encoder representations. Similarly, (150)'s TagFog, while a general visual OOD method, is motivated by applications like autonomous driving and uses textual anchor guidance from large language models (e.g., ChatGPT) and jigsaw-based fake outlier generation to train robust visual encoders. This allows for learning more compact ID representations and leaving spare regions for OOD data in the feature space, enhancing open-vocabulary OOD capabilities. The broader landscape of OOD detection in the VLM era, as surveyed by (101), underscores the transformative potential of these models for detecting novel, semantically rich OOD events that traditional methods might miss.

In conclusion, OOD detection is an indispensable enabler for safe and robust autonomous operation. Significant progress has been made in developing methods that of-

fer statistical guarantees, enhance robustness against adversarial attacks, and, critically, leverage multimodal sensor fusion for a more comprehensive understanding of the operational environment. The emergence of VLM-based approaches further promises to extend OOD capabilities to truly open-world, semantically rich unknown scenarios. However, challenges persist in developing unified theoretical frameworks that seamlessly integrate OOD detection with the broader concept of Out-of-Model-Scope, especially in highly dynamic, multimodal, and continually learning systems. Future directions should focus on scaling these guarantees to highly complex, distributed CPS, further integrating human feedback for adaptive learning, and establishing comprehensive benchmarks that reflect the full spectrum of OOD events and temporal dependencies in real-world autonomous environments, including adverse weather conditions and sensor degradation.

6.3 OOD in Cybersecurity and Anomaly Detection

Out-of-distribution (OOD) detection is a cornerstone of modern cybersecurity and anomaly detection, providing a critical defense against the dynamic and adversarial nature of digital threats. In these high-stakes environments, OOD samples frequently represent malicious activities, ranging from sophisticated network intrusions and advanced persistent threats to novel malware and fraudulent financial transactions. The ability to promptly and accurately identify these deviations from established normal patterns is paramount for safeguarding critical infrastructure, sensitive data, and financial systems. This subsection synthesizes how OOD detection methods are specifically adapted and applied to diverse data streams, including network traffic, system logs, user behavior, and graph-structured data, highlighting their utility in protecting against unpredictable and evolving threats.

A primary challenge in cybersecurity is the real-time detection of novel network intrusions and traffic anomalies amidst high-volume data streams. Early OOD methods focused on efficiency and feature-space analysis to meet these demands. For instance, Neural Mean Discrepancy (NMD) (19) offers an efficient post-hoc technique for detecting OOD samples by measuring deviations in deep neural network activation means, making it suitable for rapid monitoring of network traffic. Similarly, FeatureNorm and NormRatio

(25) identify optimal intermediate layers where in-distribution (ID) and OOD data exhibit maximal feature norm separation, providing a robust signal for unusual traffic patterns without requiring explicit OOD training samples. Addressing the need for rapid identification of new types of malicious traffic with limited labeled data, SPN (76) proposes a few-shot learning approach based on a Siamese Prototypical Network, incorporating margin loss to ensure OOD detection capabilities for unknown traffic types. In specialized network contexts, such as Controller Area Network (CAN) bus intrusion detection, a cascaded two-stage classification architecture leveraging an Auxiliary Classifier Generative Adversarial Network (ACGAN) effectively distinguishes known attacks from normal traffic while detecting unknown attack classes as OOD (26), demonstrating architectural adaptations for resource-constrained environments.

For malware analysis and system log anomaly detection, the focus shifts to distinguishing subtle, potentially polymorphic threats from benign system variations, often under conditions of data scarcity. Methods that refine internal representations are crucial here. RankFeat (5) enhances OOD detection by removing dominant rank-1 feature components that might obscure subtle OOD signals, proving effective for identifying novel threats in complex datasets like malware binaries. To mitigate ambiguity caused by atypical ID samples, Batch Normalization Assisted Typical Set Estimation (BATS) (9) rectifies extreme features to form a "typical set," which is vital for distinguishing subtle malicious anomalies in system logs from benign, yet unusual, system variations. Furthermore, MOODv2 (115) enhances ID representation learning through Masked Image Modeling (MIM), yielding more robust and distinct ID features critical for distinguishing subtle malware variants or sophisticated intrusion attempts. In unsupervised settings, where labeled anomalies are rare, Density of States Estimation (DoSE) (12) leverages multiple summary statistics from generative models to identify atypical samples, overcoming the common challenge of generative models assigning high likelihoods to OOD data, a critical consideration for unsupervised anomaly detection in logs or network flows. Beyond feature-level analysis, neuron-centric approaches like Neuron Activation Coverage (NAC) (30) quantify the "coverage degree" of neuron states to detect OOD, proving useful for

identifying deviations in learned patterns of user behavior or system states that could indicate a compromise or insider threat.

The inherently adversarial nature of cybersecurity necessitates OOD detection methods that are robust to manipulation. Attackers actively seek to bypass detectors by crafting adversarial examples that appear in-distribution. To counter this, Adversarial Learning with inlier and Outlier Exposure (ALOE) (11) trains models against both adversarial ID and OOD examples, significantly improving robustness against malicious perturbations designed to evade detection. Building on this, Adversarially Robust OOD Detection Using Lyapunov-Stabilized Embeddings (AROS) (131) leverages Neural Ordinary Differential Equations (NODEs) and Lyapunov stability to achieve robust embeddings. AROS notably generates "fake OOD embeddings" from low-likelihood regions of the ID feature space, eliminating the need for auxiliary OOD datasets and enhancing robustness against strong adversarial attacks. Gradient-based methods also contribute to robustness; GradOrth (47) identifies OOD samples by projecting gradients onto low-rank subspaces of ID data, offering a nuanced way to detect deviations in model processing. Similarly, GAIA (77) detects "abnormality" in gradient-based attribution results, providing interpretability for security analysts investigating suspicious activities. Furthermore, the concept of tangent distance (162) addresses the "curse of dimensionality" in high-dimensional feature spaces, offering a data structure-aware approach to quantify OOD uncertainty by measuring distance to the nearest submanifold space, which is crucial for robust OOD detection against subtle perturbations.

For graph-structured data, prevalent in network topology, social networks, and financial transaction graphs, OOD detection faces unique challenges due to non-Euclidean data structures and complex relationships. GOOD-D (6) pioneers unsupervised graph-level OOD detection using perturbation-free hierarchical contrastive learning. Addressing the scarcity of OOD data for graphs, GOLD (48) implicitly generates adversarial latent samples to enhance detection without auxiliary OOD datasets. GOODAT (105) offers a test-time, plug-and-play graph OOD detection method that leverages a graph masker guided by the Information Bottleneck principle, providing an efficient solution for

monitoring network intrusions. The growing importance of this domain is underscored by comprehensive surveys like (179) and unified benchmarks such as UB-GOLD (106), which allows for rigorous comparison of unsupervised graph-level anomaly and OOD detection methods across various threat scenarios. For node-level OOD detection in graph neural networks, NODESAFE (144) optimizes energy scores to reduce extreme values and mitigate logit shifts, significantly improving detection accuracy against structural manipulations.

The rise of Large Language Models (LLMs) and multimodal data streams has opened new avenues for detecting sophisticated threats like phishing and social engineering. A survey by (120) systematically reviews how LLMs are utilized for anomaly and OOD detection across various data modalities, including text. For multi-modal OOD detection, (60) leverages LLMs' world knowledge to generate descriptive features while calibrating for hallucination, enhancing the detection of complex, multi-modal threats. Envisioning Outlier Exposure (EOE) (103) utilizes LLMs to generate synthetic outlier class labels, providing "envisioned outlier exposure" to improve zero-shot OOD detection without real OOD data, which is crucial for identifying novel attack patterns or zero-day exploits. Furthermore, MIntOOD (88) proposes a multimodal intent understanding system that simultaneously achieves classification and OOD detection by fusing text, video, and audio, vital for detecting anomalous user behavior or sophisticated social engineering attacks.

In safety-critical Cyber-Physical Systems (CPS), OOD detection demands strong theoretical guarantees and real-time applicability. Inductive Conformal Anomaly Detection (ICAD) (20), using learned nonconformity measures, provides statistically sound false alarm rate guarantees for real-time OOD detection in CPS, crucial for applications like autonomous vehicles and industrial control systems. Building on this, iDECODe (17) introduces a novel non-conformity measure based on in-distribution equivariance, further strengthening conformal OOD detection with bounded false detection rates. Extending these guarantees to temporal data, CODiT (87) provides OOD detection with conformal guarantees for time-series data in CPS, directly applicable to monitoring network traffic and system logs for evolving threats. To address the practical issue of high false positive

rates in dynamic environments, a human-in-the-loop framework (118) adaptively updates OOD detection thresholds with theoretical guarantees on false positive rate control, ensuring trustworthy deployment. Furthermore, understanding the fundamental objectives of OOD methods, as explored by (42), helps in designing more principled and effective security detectors. The Model-Specific OOD framework (56) offers a unified perspective on OOD detection based on a deployed model's actual misclassification behavior, which is highly relevant for understanding what a security system *actually* fails on in a real-world context. Finally, Continual Evidential Deep Learning (CEDL) (78) integrates evidential deep learning into a continual learning framework to simultaneously perform incremental classification and OOD detection, a critical capability for systems facing evolving threats without catastrophic forgetting.

In conclusion, OOD detection is an indispensable and rapidly evolving field within cybersecurity and anomaly detection. It has progressed from basic statistical deviation measures to sophisticated, robust, and context-aware methodologies capable of addressing diverse data types and adversarial environments. While significant advancements have been made in developing methods for various data modalities, enhancing robustness against adversarial threats, and providing theoretical guarantees, several challenges persist. These include balancing the need for universal OOD solutions with the demonstrated benefits of domain-specific adaptations, developing scalable and theoretically sound methods that can handle the full spectrum of real-world distribution shifts without relying on scarce OOD training data, and rigorously aligning OOD detection with the ultimate goal of ensuring model safety and reliability in constantly changing, unpredictable digital environments. Future research must continue to bridge theoretical rigor with practical deployment, particularly in the face of increasingly sophisticated and adaptive cyber threats.

6.4 Practical Deployment Considerations and Human-in-the-Loop

The successful transition of Out-of-Distribution (OOD) detection systems from controlled experimental settings to real-world applications hinges on addressing a critical set of prac-

tical deployment challenges. Beyond theoretical performance metrics, these systems must demonstrate computational efficiency, scalability, robustness to dynamic environments, provable guarantees on false detection rates, and the capacity for effective human-in-the-loop (HITL) interaction and interpretability. The overarching goal is to develop OOD solutions that are not only technically effective but also robust, efficient, interpretable, and ultimately practical for diverse operational settings, particularly in safety-critical domains.

A foundational requirement for practical deployment, especially in latency-sensitive systems, is computational efficiency and scalability. Early OOD methods, particularly those relying on complex generative models, often incurred significant computational overhead (4). Consequently, recent research has prioritized lightweight, post-hoc approaches that minimize inference time. Neural Mean Discrepancy (NMD) (19), for instance, leverages running average means from Batch Normalization layers to approximate training data statistics, enabling real-time detection with minimal computational burden. Similarly, GradOrth (47) offers an efficient gradient-based OOD detector by pre-computing a low-rank subspace of in-distribution data gradients, facilitating rapid OOD scoring during inference. While both methods offer efficiency gains, NMD's reliance on Batch Normalization statistics might limit its applicability to architectures without such layers or in scenarios where batch statistics are highly variable. GradOrth, conversely, requires gradient computations, which, while pre-computed, still adds a layer of complexity compared to simpler confidence-based scores. The computational burden of traditional kernel methods, such as Kernel PCA (KPCA), has also been significantly reduced by approaches like CoP and CoRP (111), which devise explicit non-linear feature mappings to achieve state-of-the-art performance with $O(1)$ or $O(M)$ complexity, a substantial improvement over $O(N\_tr)$ for methods like k-Nearest Neighbors (KNN). For Cyber-Physical Systems (CPS) demanding real-time responses, (20) demonstrated efficient OOD detection by integrating learned nonconformity measures (from VAEs and Deep SVDD) into the Inductive Conformal Anomaly Detection (ICAD) framework, overcoming traditional conformal prediction's scalability limitations for high-dimensional sensor inputs. Furthermore, the

efficiency challenge extends to large language models (LLMs) in natural language processing (NLP). (79) proposed PTO, an unsupervised prefix-tuning based OOD detection framework that offers a parameter-efficient alternative to costly fine-tuning, demonstrating comparable or superior performance with significantly reduced storage and computational requirements. These diverse approaches highlight a collective effort to balance detection efficacy with the stringent computational constraints of real-world deployment.

Beyond raw efficiency, practical systems demand robustness to diverse OOD types and adversarial attacks, coupled with adaptive mechanisms for dynamic, non-stationary environments. Many methods improve intrinsic OOD robustness during training or representation learning (as discussed in Section 4), but deployment-time robustness necessitates adapting to unforeseen shifts. The conceptual shift from merely detecting "out-of-distribution" to "Out-of-Model-Scope" (OMS) detection (29) is crucial, emphasizing the identification of inputs leading to *prediction errors* of the specific deployed model, rather than a generic notion of novelty. This perspective is further refined by the Model-Specific Out-of-Distribution (MS-OOD) framework (56), which unifies the detection of semantic shifts, covariate shifts, and even misclassified in-distribution examples based on the actual performance of a deployed classifier. This framework is vital for guiding adaptive behavior and dynamic thresholding, allowing systems to differentiate between inputs the model *can* handle despite a shift (e.g., a covariate shift it generalizes to) and those it *cannot* (e.g., a semantic shift or a covariate shift leading to misclassification). While MS-OOD provides a robust conceptual foundation, its practical implementation for continuous adaptation in dynamic environments remains an active area of research, often requiring continuous monitoring and re-calibration. Contributions like (174)'s sparsity-regularized tuning framework enhance the generalizability of OOD score functions, making them less dependent on specific datasets and more capable of dynamic adaptation. A particularly innovative adaptive mechanism for handling unforeseen OOD in open-world scenarios is (103)'s Envisioning Outlier Exposure (EOE). EOE leverages Large Language Models (LLMs) to synthetically generate potential outlier class labels based on visual similarity to in-distribution classes, effectively providing "outlier exposure" without requiring ac-

tual OOD data. This LLM-driven approach offers a scalable and flexible way to adapt to various OOD types (far, near, fine-grained) by dynamically envisioning new categories, thereby enhancing the model's ability to distinguish novel inputs in a zero-shot manner. However, the effectiveness of EOE relies heavily on the quality of LLM-generated prompts and the LLM's inherent knowledge, posing challenges for robust and unbiased outlier generation across all domains.

A critical aspect of practical deployment, especially in safety-critical domains, is the ability to provide reliable uncertainty estimates and control false detection rates. This is paramount for building trust and ensuring regulatory compliance. Conformal Prediction (CP), as detailed in Section 7.2, offers a principled approach to this, providing provably valid false detection rates. For instance, (17) introduced iDECODe for single-point OOD detection with theoretically guaranteed bounded False Detection Rates (FDR). This work was significantly extended by (87) to time-series data in Cyber-Physical Systems, providing conformal guarantees for OOD detection in dynamic, dependent data streams. This is a crucial advancement, as many real-world applications involve sequential data where independence assumptions of traditional CP might not hold. While CP offers strong theoretical guarantees, its practical application often requires careful calibration and consideration of the exchangeability assumption, which can be challenging in highly non-stationary environments.

Despite theoretical guarantees, managing false positives (FPs) in dynamic, open-world settings often requires human oversight, leading to the indispensable role of human-in-the-loop (HITL) approaches. HITL frameworks integrate human expert feedback to refine OOD detectors, manage trade-offs between safety and performance, and build trust. (118) directly addressed the problem of high False Positive Rates (FPR) in OOD detection by proposing a mathematically grounded HITL framework. This framework adaptively updates the detection threshold over time by integrating human feedback and employing an anytime-valid Upper Confidence Bound (UCB) based on the Law of Iterated Logarithm, guaranteeing FPR control below a desired level while maximizing True Positive Rate (TPR) (further theoretical details are in Section 7.2). This approach offers a ro-

bust mechanism for dynamic threshold adjustment, but its effectiveness depends on the availability and reliability of human feedback. Beyond direct threshold adjustment, HITL also plays a crucial role in data acquisition and model refinement. (89) proposed SISOM, a unified approach for active learning and OOD detection. Active learning inherently involves human experts labeling uncertain or novel samples, thereby providing crucial feedback to improve both model performance and OOD detection capabilities. SISOM's self-deciding process for combining scores contributes to adaptive behavior, reducing the burden on human operators while maintaining robustness.

For HITL systems to be truly effective, building operator trust and ensuring practical utility requires OOD solutions that are not only effective but also interpretable and aligned with system safety goals. Human operators need to understand *why* a system flags an input as OOD to make informed decisions and foster confidence in AI systems. Methods that leverage intrinsic model properties or explanations can enhance this interpretability. For instance, Neuron Activation Coverage (NAC) (30) quantifies the "coverage degree" of neuron states by in-distribution data, offering insights into model behavior under OOD conditions. Similarly, GAIA (77) detects OOD samples by quantifying "abnormality in gradient-based attribution results," directly linking model explanations to OOD detection (these methods are detailed in Section 4.3). While these approaches provide valuable diagnostic information, translating complex neural network activations or gradient attributions into easily digestible and actionable insights for human operators remains a significant challenge. The interpretability must be tailored to the human's cognitive load and the specific decision-making context.

In conclusion, the literature demonstrates a clear trajectory towards OOD detection solutions that prioritize practical deployment. This involves a strong emphasis on computational efficiency for real-time operation, robustness and adaptive mechanisms for dynamic environments, and the provision of theoretical guarantees on false alarm rates. Crucially, the field is increasingly recognizing the indispensable role of human-in-the-loop approaches for adaptive thresholding, managing false positives, and building trust in AI systems. Future work will likely continue to explore more nuanced human-AI collabora-

tion models, develop methods for handling highly dynamic and unforeseen OOD shifts with minimal human intervention, and strive for truly interpretable OOD explanations that align with human decision-making in safety-critical contexts, ultimately enabling the responsible and reliable deployment of AI.

7 Ensuring Trustworthy OOD: Advanced Formalisms, Guarantees, and Evaluation

7.1 Evolving OOD Definitions and Granular Taxonomies

The conceptualization of Out-of-Distribution (OOD) detection has undergone a significant evolution, moving beyond a simplistic binary distinction between in-distribution (ID) and OOD data towards a more nuanced, granular, and context-aware understanding. This shift is critical for developing robust and trustworthy AI systems capable of operating reliably in complex real-world environments.

Initially, OOD detection primarily focused on identifying samples from entirely novel semantic categories. However, this narrow view proved insufficient, leading to the introduction of more granular definitions. A pivotal development was the formal distinction between different types of distribution shifts. (3) introduced the concept of *Full-Spectrum Out-of-Distribution (FS-OOD) Detection*, explicitly differentiating between *semantic shift* (novel classes) and *covariate shift* (label-preserving appearance changes like lighting or style). Their proposed Semantics score function (SEM) aimed to disentangle these shifts, demonstrating that existing methods often failed to robustly handle covariate-shifted ID data, treating it erroneously as OOD. Further complicating the landscape, (8) highlighted the critical impact of spurious correlations, formalizing "spurious OOD" where models rely on non-causal features, making detection challenging even for inputs that visually resemble ID data. This emphasized that OOD can arise not just from novel semantics or appearance, but also from the model's learned biases.

The need for more rigorous evaluation of these granular shifts quickly became apparent.

(54) critically analyzed existing ImageNet-based OOD benchmarks, revealing issues such as ID contamination, semantic ambiguities, and unintended covariate shifts that hindered the accurate assessment of semantic OOD detection. To address this, they introduced `ImageNet-OOD`, a meticulously human-curated dataset designed to isolate pure semantic shift by minimizing covariate variations. Building on this, (140) provided a comprehensive cross-evaluation of OOD detection and Open-Set Recognition (OSR) methods, further disentangling semantic and covariate shifts on large-scale benchmarks and proposing a new "Outlier-Aware Accuracy" (OAA) metric to reconcile robustness to covariate shift with the ability to detect its presence. These works collectively underscored the importance of clean, disentangled evaluation for understanding what OOD algorithms truly detect.

A significant conceptual re-framing of the OOD problem was introduced by (56) with the *Model-Specific Out-of-Distribution (MS-OOD) Detection* framework. This paradigm shifted the definition of OOD from being purely based on data properties to being dependent on a *specific deployed model's performance and misclassification behavior*. Under MS-OOD, an example is considered OOD if the model cannot classify it correctly, unifying semantic shift, covariate shift, and even misclassified in-distribution examples under a single, performance-driven ground truth. This perspective acknowledges that what constitutes "OOD" can be subjective and model-dependent, moving towards a more practical, context-aware definition.

The binary nature of traditional OOD evaluation also faced scrutiny. (178) addressed the "Sorites Paradox" in OOD evaluation, arguing that a simple binary ID/OOD distinction fails to capture the continuous *degree* of semantic and covariate shifts. They introduced the *Incremental Shift OOD (IS-OOD)* benchmark and *Language Aligned Image feature Decomposition (LAID)*, a CLIP-based method to quantitatively decompose image features into distinct semantic and covariate components. This allowed for continuous measurement of shift degrees, providing a far more granular and informative evaluation of OOD detection performance as a function of shift intensity.

As the field matured and its problem definitions became increasingly complex, the need for structured organization emerged. (24) provided the first comprehensive survey

on OOD detection in Natural Language Processing (NLP), introducing a novel taxonomy based on the availability of OOD data during training and highlighting NLP-specific challenges. This reflects the emergence of task-oriented and domain-specific taxonomies to organize the field's growing complexity, moving beyond generic definitions to practical, application-driven categorizations. Complementing this, theoretical works like (121) and (126) contribute to this maturing conceptual understanding by exploring the fundamental learnability of OOD detection and the role of in-distribution labels, implicitly influencing how OOD boundaries are conceptualized and defined under various conditions.

In conclusion, the evolution of OOD definitions has progressed from a rudimentary binary classification to a sophisticated, multi-faceted understanding. This trajectory, marked by the differentiation of semantic and covariate shifts, the adoption of model-specific perspectives, the development of continuous shift measurements, and the emergence of task-oriented taxonomies, collectively reflects a maturing conceptual understanding of OOD detection. However, challenges remain in developing scalable, universally applicable methods that can robustly handle the full spectrum of these granular shifts without relying on scarce OOD training data, particularly in diverse real-world applications.

7.2 Certifiable OOD Detection: Provable Guarantees and Conformal Prediction

The deployment of machine learning systems in safety-critical applications necessitates not only high empirical performance but also strong theoretical guarantees regarding their reliability, particularly in identifying out-of-distribution (OOD) inputs. This subsection explores the critical advancements towards certifiable OOD detection, emphasizing methods that provide provable guarantees on false detection rates and leverage rigorous statistical frameworks like Conformal Prediction (CP).

A cornerstone of certifiable OOD detection is the integration of Conformal Prediction (CP), which offers a robust, model-agnostic framework for controlling false detection rates with statistical validity. Early work by (20) introduced Inductive Conformal Anomaly De-

tection (ICAD) for real-time OOD detection in learning-enabled Cyber-Physical Systems (CPS). This approach overcame the scalability limitations of traditional CP by employing learned nonconformity measures (NCMs) based on Variational Autoencoders (VAEs) and Deep Support Vector Data Description (SVDD), ensuring a well-calibrated false alarm rate in high-dimensional settings. Building on this, (17) proposed iDECODe, a novel ICAD framework leveraging in-distribution equivariance as its NCM. By aggregating scores from multiple transformations, iDECODe provides a theoretically guaranteed bounded false detection rate, demonstrating state-of-the-art performance in single-point OOD detection. Extending these guarantees to dynamic environments, (87) developed CODiT, which applies conformal anomaly detection to dependent time-series data in CPS. CODiT uses deviation from in-distribution temporal equivariance as an NCM and combines predictions from multiple detectors via Fisher's method, offering bounded false alarms for both fixed-length windows and variable-length traces. These advancements collectively showcase CP's versatility in providing marginal coverage guarantees across diverse OOD scenarios, from static single-point detection to complex temporal data streams.

Beyond merely providing detection guarantees, CP is also crucial for establishing statistically rigorous evaluation metrics for OOD detectors themselves. Traditional OOD evaluation metrics, such as AUROC and FPR@TPR95, are empirical approximations that can be overly optimistic and fluctuate significantly with finite test sample sizes, lacking robust, conservative guarantees. Addressing this, (149) proposed a dual application of CP and OOD detection. They introduced "conformal AUROC" and "conformal FPR" metrics, which provide probabilistic conservativeness guarantees on the variability of these evaluation metrics. This ensures that the estimated performance of an OOD detector is conservative with high probability (e.g., $1-\delta$), thereby making the *evaluation* of OOD systems certifiable and more trustworthy. Furthermore, (149) demonstrated that sophisticated OOD scores, such as Mahalanobis distance or K-Nearest Neighbors (KNN) distance, can serve as highly effective non-conformity scores within the CP framework, often outperforming classical CP non-conformity scores in building prediction sets. This highlights a synergistic relationship where OOD methods can enhance CP, and CP can,

in turn, provide robust evaluation for OOD.

While CP provides statistical guarantees, real-world systems often require adaptive control and human oversight, especially when faced with evolving OOD distributions. Addressing this, (118) introduced a mathematically grounded human-in-the-loop framework for OOD detection that dynamically adjusts detection thresholds. This framework leverages importance sampling and an anytime-valid Upper Confidence Bound (UCB) based on the Law of Iterated Logarithm to provide provable guarantees on the false positive rate (FPR), even in the presence of distribution shifts. By taming false positives with minimal human feedback, this approach significantly enhances the practical deployability and trustworthiness of OOD systems in dynamic, open-world environments.

Beyond statistical guarantees and adaptive control, a deeper theoretical understanding of OOD detection learnability and data utility is crucial for a trustworthy foundation. (95) made a significant contribution by providing the first framework that offers *provable guarantees* for leveraging unlabeled "wild" data in OOD detection. Their "Separate And Learn" (SAL) framework employs a novel gradient-based filtering mechanism and offers rigorous error bounds on outlier separability and classifier learnability, demonstrating how unlabeled data can provably enhance OOD awareness without requiring clean auxiliary OOD samples. Complementing this, (126) delved into the fundamental learnability of OOD detection, establishing necessary and sufficient conditions for Probably Approximately Correct (PAC) learnability under various risk and AUC metrics. This theoretical work highlights that OOD detection is not universally learnable and depends critically on the characteristics of the data and hypothesis spaces, providing crucial insights into the theoretical limits and possibilities of certifiable OOD systems. The implications for CP are profound: while CP offers statistical guarantees *given* an OOD score, the inherent quality and effectiveness of that score, and thus the practical utility of the CP-based detection, are constrained by the underlying learnability conditions of the specific OOD problem. This underscores the need for OOD scores that are well-aligned with the learnable properties of the data distribution for CP to be truly effective in practice.

In conclusion, the pursuit of certifiable OOD detection is rapidly evolving, moving

from foundational statistical guarantees offered by Conformal Prediction for detection and evaluation, to adaptive, human-in-the-loop frameworks for dynamic FPR control. Simultaneously, theoretical work is establishing the learnability and data utility principles, collectively shaping a more robust and trustworthy understanding of what it means for an AI system to be "certifiably" aware of its own limitations. Despite these advancements, challenges remain in developing universally applicable nonconformity measures for CP that can provide strong guarantees across all types of distribution shifts, scaling CP to increasingly large foundation models, and extending these guarantees to more complex adaptive or continually learning systems, all while ensuring that the underlying OOD problem is indeed theoretically learnable.

7.3 Standardized Benchmarking and Unified Evaluation Frameworks

The systematic and fair advancement of Out-of-Distribution (OOD) detection critically relies on the development of standardized benchmarks and unified evaluation frameworks. Historically, the field grappled with inconsistent definitions of OOD, ad-hoc datasets, and disparate evaluation protocols, severely hindering reproducible and comparable research. Early work by (8) highlighted this by formalizing the concept of "spurious OOD," demonstrating how models' reliance on spurious correlations in training data could lead to high-confidence but unreliable predictions on OOD inputs, thus exposing a fundamental flaw in simplistic OOD definitions and evaluation. Similarly, (31)'s comparative study revealed significant performance discrepancies of confidence-based OOD methods between general computer vision tasks and challenging medical imaging applications, underscoring the necessity for domain-specific and rigorous evaluation. The lack of a comprehensive review for OOD in Natural Language Processing (NLP) was addressed by (24), which provided a taxonomy and discussed NLP-specific evaluation challenges, while (108) offered a systematic framework and taxonomy for OOD detection in medical image analysis, clarifying terminology and evaluation protocols for this critical domain. These initial efforts underscored the urgent need for a more structured and consistent approach to OOD evaluation.

A pivotal development in OOD evaluation has been the effort to disentangle distinct types of distribution shifts, moving beyond a monolithic view of OOD (as discussed in detail in Section 7.1). To operationalize these nuanced definitions, researchers have meticulously curated datasets designed to isolate and evaluate performance on specific shifts. (54) meticulously curated *ImageNet-OOD*, a novel dataset specifically designed to isolate and evaluate performance on semantic shifts while minimizing confounding covariate shifts. This dataset addressed critical shortcomings of previous ImageNet-based benchmarks, such as ID contamination and semantic ambiguities. Its findings were impactful, demonstrating that many modern OOD algorithms are disproportionately sensitive to covariate shifts rather than genuine semantic novelty, often failing to detect truly novel classes. This revelation prompted a re-evaluation of existing methods and guided the development of more robust techniques. Furthering this disentanglement, (140) provided a critical analysis of OOD detection and Open-Set Recognition (OSR) methods, introducing a new large-scale benchmark to systematically disentangle semantic and covariate shifts and proposing "Outlier-Aware Accuracy" as a more nuanced metric. Complementing these efforts, (178) introduced the "Incremental Shift OOD" (IS-OOD) benchmark, which categorizes OOD samples by their *degree* of semantic and covariate shift, moving beyond binary OOD definitions and utilizing a novel Language Aligned Image feature Decomposition (LAID) method to quantify these shifts, offering a more granular assessment of OOD robustness.

Beyond specialized datasets, the field has seen the emergence of comprehensive, unified software frameworks that provide robust platforms for rigorous comparison across diverse methods and OOD scenarios, addressing the critical need for reproducible and scientifically sound assessments. A cornerstone in this regard is **OOD-Bench** (**?**), which emerged to tackle the pervasive issues of inconsistent implementations and evaluation settings. OOD-Bench provides a modular and extensible platform that integrates a wide array of OOD detection methods, backbone architectures, and datasets, enabling researchers to conduct fair and reproducible comparisons. Its structured approach has significantly improved the reliability of reported results and fostered a more systematic

advancement of the field. Similarly, (23) introduced **PyTorch-OOD**, a Python library specifically designed to accelerate OOD detection research and improve reproducibility. By providing well-tested and documented implementations of OOD methods with a unified interface, along with benchmark datasets and utility functions, PyTorch-OOD lowers the barrier to entry for new researchers and ensures consistency across experiments.

The demand for standardized evaluation extends to diverse data modalities and complex learning scenarios. For graph-structured data, which presents unique challenges due to its non-Euclidean nature, (6) pioneered a benchmark dataset for unsupervised graph-level OOD detection. This was significantly expanded by (106) with **UB-GOLD**, a unified benchmark that integrates unsupervised graph-level anomaly detection and OOD detection across 35 datasets and four distinct scenarios. UB-GOLD provides a robust and comprehensive platform for rigorous comparison in this emerging domain, allowing for a deeper understanding of method performance under various graph OOD conditions. A recent survey by (179) further contributes to the standardization of graph OOD detection by providing a rigorous definition and systematically categorizing existing methods, clarifying distinctions with related fields and highlighting unique challenges. In medical imaging, where OOD detection is paramount for patient safety, (64) developed a novel and diverse benchmark for 3D medical image segmentation OOD, which exposed the significant limitations of many state-of-the-art methods when confronted with the subtle, yet critical, OOD shifts inherent in medical data. Similarly, (65) contributed a new benchmark for OOD detection in medical imaging by manually annotating pacemakers and support devices in chest X-rays, enabling more targeted evaluation of methods like Mahalanobis distance against clinically relevant anomalies.

Furthermore, as OOD detection integrates with more dynamic learning paradigms, specialized benchmarks become crucial. **OpenCIL** (158) addresses the critical challenge of OOD detection within Class-Incremental Learning (CIL). CIL models, designed to continuously learn new classes, often suffer from catastrophic forgetting, which severely impacts their ability to detect OOD samples reliably. OpenCIL is the first comprehensive benchmark for OOD detection in CIL, providing unified evaluation protocols and

two principled frameworks (post-hoc and fine-tuning based) to integrate OOD methods into CIL models. This benchmark has been instrumental in identifying critical biases in CIL models towards OOD samples and newly added classes, offering crucial insights for designing future open-world CIL systems.

In conclusion, the evolution of OOD detection has seen a critical shift from ad-hoc evaluations to sophisticated, standardized benchmarking and unified evaluation frameworks. The development of meticulously curated datasets like *ImageNet-OOD* (54) has been indispensable for disentangling various types of distribution shifts, while comprehensive software platforms such as OOD-Bench (**?**) and PyTorch-OOD (23) provide the necessary infrastructure for reproducible and fair comparisons. Specialized benchmarks like UB-GOLD (106) for graph data, medical imaging benchmarks (64; 65), and OpenCIL (158) for CIL have expanded the field's evaluative rigor into complex domains. Despite these advancements, challenges remain in creating truly exhaustive benchmarks that capture the full spectrum of real-world OOD scenarios, particularly for dynamic and 'near-OOD' shifts, and in developing evaluation protocols that scale effectively for increasingly large foundation models. Future research must continue to bridge the gap between empirical performance and theoretical understanding, ensuring that benchmarks not only measure but also drive the development of truly robust and trustworthy OOD solutions.

8 Conclusion and Future Directions

8.1 Synthesis of Key Trends and Contributions

The field of Out-of-Distribution (OOD) detection has undergone a profound transformation, evolving from rudimentary post-hoc scoring mechanisms to sophisticated, context-aware strategies that leverage advanced model architectures and rigorous theoretical foundations. This progression is driven by the imperative to build more reliable, adaptable, and trustworthy AI systems capable of operating effectively and safely in unpredictable, open-world environments. The collective research consolidates our understanding of how

OOD detection has matured to address the growing demands for robust uncertainty quantification across diverse applications.

Initially, research focused on extracting OOD signals from already trained models, often through feature engineering and statistical analysis. Early efforts explored the utility of reconstruction-based methods, with (2) rethinking autoencoder-based OOD by introducing layerwise semantic reconstruction and a Normalized L2 Distance to make reconstruction error a more valid uncertainty measure. Complementing this, (7) proposed embedding in-distribution (ID) data into a union of 1-dimensional subspaces for compact representation and easier OOD detection. The analysis of feature properties also proved fruitful: (5) introduced RankFeat, a post-hoc method that removes a dominant rank-1 component from high-level features based on spectral analysis, significantly improving performance. Similarly, (9) boosted OOD detection by rectifying features into their "typical set" using a Truncated Batch Normalization unit, mitigating the impact of extreme features. (25) further explored feature norms, demonstrating that intermediate layers often provide better OOD separation than the final layer, and proposed a block selection method using pseudo OOD data. Challenging the prevailing reliance on simple output-based scores, (38) revisited OOD baselines and strongly advocated for the effectiveness of k-Nearest Neighbor (KNN) distance on learned embeddings. More recently, (93) advanced distance-based methods by modeling ID classes with a mixture of prototypes in a hyperspherical embedding space, capturing intra-class diversity, while (111) demonstrated the power of Kernel PCA with efficient explicit feature mappings for non-linear OOD separation. These methods collectively refined the ability to discern OOD samples from subtle cues within a model's internal representations, often without requiring additional training.

A significant intellectual trajectory involved moving beyond passive post-hoc analysis to actively enhancing model robustness during training. This included adversarial training, as seen in (11)'s ALOE, which robustified OOD detectors against both adversarial in-distribution and OOD examples. A major paradigm shift was the widespread adoption and refinement of Outlier Exposure (OE), where auxiliary OOD data is used to regularize model training. (18) introduced Mixture Outlier Exposure (MixOE) to address

fine-grained OOD by mixing ID and auxiliary data, creating a broader virtual outlier distribution. Providing theoretical grounding, (42) demonstrated that many OE methods are asymptotically equivalent to a binary discriminator, highlighting that differences often stem from estimation procedures. Subsequent work focused on optimizing the utility of auxiliary data: (72) proposed Diverse Outlier Sampling (DOS) to select diverse and informative outliers, a concept further advanced by (85)'s diverseMix, which provably enhances outlier diversity through semantic-level interpolation. Addressing practical challenges, (39) introduced a balanced energy regularization loss to account for class imbalance within auxiliary OOD data, while (116) leveraged Energy-based Hopfield Boosting for adaptive sampling of "hard" outliers. Complementing these data-centric strategies, methods like (63)'s Average of Pruning (AoP) tackled training instability and overfitting in OOD detection, a theme further explored by (86) with optimal parameter and neuron pruning based on gradient sensitivity. (84) explicitly pursued feature separation based on Neural Collapse, constraining OOD features to an orthogonal subspace of ID features during fine-tuning. Concurrently, generative models also evolved: (4) introduced Deep Residual Flow for improved density modeling in feature activations, and (12)'s Density of States Estimation (DoSE) shifted focus from direct likelihoods to the typicality of multiple summary statistics, overcoming the "high likelihood for OOD" pathology.

The field has also expanded dramatically to encompass complex data modalities, specialized learning paradigms, and the formidable capabilities of foundation models. For dense prediction tasks, (15) proposed Residual Pattern Learning (RPL) for pixel-wise OOD detection in semantic segmentation, decoupling it from the main task. (28) further enhanced segmentation OOD by learning from local adversarial attacks to generate OOD-like training data, while (73)'s ATTA introduced anomaly-aware test-time adaptation to handle domain shifts. For graph-structured data, (6) pioneered unsupervised graph-level OOD detection with GOOD-D, a hierarchical contrastive learning framework, a direction further advanced by (48)'s GOLD, which uses implicit adversarial latent generation to synthesize OOD samples without auxiliary data, and (105)'s GOODAT for test-time graph OOD detection. The rise of Vision-Language Models (VLMs) and Large Language

Models (LLMs) has opened new frontiers: (41) introduced GL-MCM for zero-shot OOD detection by combining global and local CLIP features, while (96) learned transferable negative prompts for open-vocabulary OOD, and (109) proposed Self-Calibrated Tuning to mitigate spurious OOD features in VLMs. Leveraging LLMs further, (60) explored their world knowledge for multimodal OOD, carefully calibrating for hallucination, and (103) used LLMs for "envisioned outlier exposure" in zero-shot settings. Multimodal OOD itself gained a dedicated benchmark with (100)'s MultiOOD, which also proposed the Agree-to-Disagree (A2D) algorithm to amplify inter-modal prediction discrepancies. This was complemented by (113)'s DPU, addressing intra-class variability in multimodal OOD. Diffusion models also found their niche, with (49) applying Latent Diffusion Models for unsupervised 3D medical OOD detection, and (44)'s DiffGuard using pre-trained diffusion models for semantic mismatch-guided OOD. The field also expanded to long-tailed recognition (35; 46), LiDAR-based 3D object detection (81), and even mathematical reasoning in GLMs using embedding trajectories (107).

Crucially, the field has placed an increasing emphasis on theoretical guarantees, robust evaluation, and a critical re-evaluation of fundamental definitions to build truly trustworthy AI. (3) introduced the Full-Spectrum OOD (FS-OOD) problem, distinguishing between semantic and covariate shifts, and proposed the SEM score for robust detection. This was followed by rigorous benchmarking efforts: (16) established the MOOD 2020 benchmark for medical imaging, (54) created ImageNet-OOD to disentangle semantic and covariate shifts, and (140) critically dissected OOD and Open-Set Recognition (OSR) methods and benchmarks. The "Sorites Paradox" in OOD evaluation was addressed by (178), proposing the Incremental Shift OOD (IS-OOD) benchmark to categorize samples by continuous shift degrees. Theoretical underpinnings have also solidified: (51) provided a principled explanation for why feature norm helps OOD detection, linking it to hidden classifier confidence, while (121) formally analyzed when and how in-distribution labels provably help OOD detection. (126) investigated the fundamental learnability of OOD detection, establishing necessary and sufficient conditions. For safety-critical systems, the focus shifted to provable guarantees: (20) developed real-time OOD detection for

Cyber-Physical Systems (CPS) with conformal guarantees using VAEs and Deep SVDD, a concept extended by (17)'s iDECODe, which leveraged in-distribution equivariance. (87) further extended this to dependent data in CPS with temporal equivariance. Critically, (29) argued that "OOD detection is not all you need," proposing Out-of-Model-Scope (OMS) detection as a more direct goal for identifying model errors, and (118) introduced a human-in-the-loop framework to tame false positives with theoretical FPR guarantees. These interconnected developments highlight a maturation of the field, moving towards comprehensive solutions that are not only performant but also interpretable, reliable, and adaptable to the complex demands of real-world AI deployment.

8.2 Open Challenges and Future Research Avenues

The quest for truly robust and autonomous AI systems hinges critically on their ability to reliably detect and appropriately handle Out-of-Distribution (OOD) inputs. Despite significant advancements, the field of OOD detection continues to grapple with several profound challenges that define current research frontiers and pave the way for future innovation.

One persistent challenge is the "near OOD" problem, where subtle shifts in data distribution are difficult to distinguish from in-distribution (ID) variations. Models often exhibit overconfidence on these semantically similar, yet novel, inputs, leading to unreliable predictions (8). Addressing this requires a multi-faceted approach. Some research focuses on **data-centric strategies** to refine the ID/OOD boundary during training. For instance, Mixture Outlier Exposure (MixOE) (18) generates virtual outliers by mixing ID and auxiliary data, specifically targeting fine-grained OOD detection where samples share visual similarities with ID data. Similarly, Virtual Outlier Smoothing (VOSo) (90) constructs virtual outliers by perturbing semantic regions of ID samples, aiming to create smoother, more robust decision boundaries. However, a key challenge remains in generating truly representative and diverse near-OOD samples without inadvertently corrupting the ID manifold. Other efforts concentrate on **representation-centric enhancements**, aiming to improve the inherent separability of ID and OOD features. Batch Normaliza-

tion Assisted Typical Set Estimation (BATS) (9) rectifies extreme features, while Variational Rectified Activation (VRA) (62) proposes optimal activation functions to improve ID/OOD separability. Leveraging Neural Collapse properties, such as ID/OOD Orthogonality (NC5) (27), projects features onto principal component spaces for better OOD detection, inherently aiding in distinguishing subtle shifts (as discussed in Section 4.2). Neuron Activation Coverage (NAC) (30) provides a novel uncertainty measure sensitive to abnormal activation patterns caused by subtle OOD inputs by quantifying neuron behavior. From a theoretical perspective, (121) highlights the crucial role of ID labels in these near-OOD scenarios. Despite these advancements, a fundamental understanding of *what constitutes a "near OOD" boundary* and how to robustly generalize detection across diverse, subtly shifted domains remains an open problem, necessitating more robust benchmarks like ImageNet-OOD (54) and IS-OOD (178) that disentangle semantic and covariate shifts for accurate evaluation.

Another significant open challenge is the scalability of OOD detection methods, particularly for increasingly large foundation models like Vision-Language Models (VLMs) and Large Language Models (LLMs). While these models offer unprecedented representational power and open-vocabulary capabilities (as explored in Section 5.3), traditional OOD methods often struggle with their computational and data demands, or fail to leverage their rich representations effectively without prohibitive inference costs or extensive fine-tuning. Current research is making strides in adapting OOD detection to this new paradigm. Approaches include leveraging pre-trained features, such as GL-MCM (41) which combines global and local CLIP features for zero-shot OOD detection, offering flexibility for multi-object scenes. **Prompt engineering and virtual outlier generation** are also emerging as scalable solutions: Outlier Label Exposure (OLE) (82) uses auxiliary outlier class labels as pseudo OOD text prompts for VLMs, and NegPrompt (96) learns transferable negative prompts from ID data alone to enhance OOD sensitivity without external outlier data. Self-Calibrated Tuning (SCT) (109) adaptively adjusts ID classification and OOD regularization in VLMs to mitigate spurious OOD features. The potential of LLMs for generating synthetic outlier exposure is explored by (103), envisioning how

LLM knowledge can create diverse outlier labels for zero-shot OOD detection. For multimodal foundation models, the MultiOOD benchmark and the Agree-to-Disagree (A2D) algorithm (100) leverage inter-modal prediction discrepancies, while Dynamic Prototype Updating (DPU) (113) accounts for intra-class variability. However, the core challenge lies in developing OOD detection frameworks that are *inherently* scalable, efficient, and robust for models with billions of parameters, without requiring extensive retraining or sacrificing the model's generalizability. This includes tackling prohibitive inference costs, catastrophic forgetting during OOD-specific fine-tuning, and the theoretical understanding of OOD behavior in these complex architectures, as highlighted by (101).

Looking ahead, future research avenues are poised to develop more adaptive and dynamic OOD systems that can learn and adjust in real-time. This involves moving beyond static OOD detectors to systems capable of continuous monitoring and adaptation in dynamic environments, such as Cyber-Physical Systems (CPS). Building on the practical deployment considerations discussed in Section 6.4, methods like those pioneered by (20) use learned nonconformity measures within a conformal prediction framework to provide real-time OOD detection with statistical guarantees. Further advancements like iDECODe (17) leverage in-distribution equivariance for conformal OOD detection with bounded false detection rates, a concept extended to dependent time-series data in (87). The challenge of adapting to domain shifts at test time for dense OOD detection in segmentation is addressed by ATTA (73), which uses a dual-level adaptation framework. Future work needs to focus on **online OOD detection** that can continuously update its model of ID and OOD without full retraining, **proactive adaptation** that anticipates shifts, and **self-correcting AI systems** that can not only detect OOD but also intelligently propose mitigation strategies or request human intervention (118). The concept of adaptive sampling of "hard" outliers during training, as demonstrated by Energy-based Hopfield Boosting (116), also contributes to dynamic system adjustment.

Another promising direction involves exploring **causal inference for OOD detection** to understand underlying mechanisms rather than merely identifying statistical anomalies. Traditional OOD methods often rely on statistical correlations, making them

vulnerable to spurious associations that do not generalize across different environments. The work by (8) on the impact of spurious correlation for OOD detection underscores this limitation. Future research should focus on developing OOD detectors that are robust to such correlations by explicitly learning causal relationships. This could involve leveraging frameworks like Invariant Risk Minimization (IRM) (**?**) or Structural Causal Models (SCMs) to disentangle causal (invariant) features from non-causal (environmental) ones. Specific research questions include: How can we design training objectives that promote the learning of causally invariant representations that are inherently more robust to OOD shifts? Can interventional or counterfactual reasoning be used to identify features that truly *cause* an input to be OOD, leading to more interpretable and generalizable OOD signals? Furthermore, exploring causal discovery techniques to model the underlying causal graph of ID data could enable the detection of OOD samples as deviations from this fundamental structure, offering a deeper, more principled understanding of novelty.

Finally, fostering deeper integration with other machine learning tasks like active learning and continual learning is crucial for holistic, efficient, and robust AI solutions that can operate autonomously in complex environments. OOD detection naturally complements **active learning (AL)**, as OOD samples represent regions of uncertainty where the model's competence is low, making them ideal candidates for human labeling. SISOM (89) proposes a unified approach, demonstrating that OOD detection and AL can be addressed simultaneously by leveraging enriched feature space distance metrics. Future work could explore how OOD uncertainty can more effectively guide AL to discover truly novel classes or subtle shifts, rather than just ambiguous ID samples. Similarly, in **continual learning (CL)**, OOD detection is vital for maintaining robustness to previously learned ID data while reliably identifying novel inputs without catastrophic forgetting. Continual Evidential Deep Learning (CEDL) (78) offers a solution for simultaneous incremental object classification and OOD detection. MIntOOD (88) extends this to multimodal intent understanding. The challenge lies in developing OOD detectors that can dynamically evolve with the model in CL settings, updating their ID boundaries without re-exposing to all past data or confusing new ID classes with true OOD. These integrated approaches

represent a significant step towards building AI systems that are not only aware of their limitations but can also actively learn, adapt, and operate safely in dynamic, open-world settings.

Ultimately, the future of OOD detection lies in a paradigm shift from reactive, isolated detectors to proactive, integrated, and self-monitoring AI systems. This grand vision entails models that continuously learn their own competence boundaries, adapt dynamically to evolving environments, leverage causal understanding for robust generalization, and seamlessly integrate with learning processes like active and continual learning. Such holistic, efficient, and robust AI solutions will be indispensable for building trustworthy systems that can operate autonomously and ethically in an increasingly complex and unpredictable world.

8.3 Ethical Considerations and Societal Impact

The integration of Out-of-Distribution (OOD) detection mechanisms into real-world AI systems, particularly in high-stakes applications, necessitates a rigorous examination of their ethical implications and potential societal impacts. Failures in OOD detection can precipitate profound consequences, ranging from critical safety hazards in autonomous systems to the perpetuation of discriminatory outcomes in sensitive decision-making processes. Consequently, advancements in this domain must transcend mere technical performance, actively embedding principles of transparency, fairness, and accountability to foster responsible and human-centric AI deployment.

A paramount ethical concern centers on the deployment of AI models in safety-critical Cyber-Physical Systems (CPS), where OOD failures can be catastrophic. For instance, autonomous vehicles and medical diagnostic tools rely heavily on robust OOD detection to prevent misinterpretations of novel inputs that could lead to severe accidents or incorrect diagnoses (20). The ethical imperative here is to ensure not only high detection rates but also controlled error rates, particularly false positives and false negatives. While the technical details of certifiable OOD detection are elaborated in Section 7.2, it is ethically crucial that such systems provide statistically bounded false detection rates, as proposed

by frameworks like conformal prediction (17; 87). These guarantees are vital safeguards against erroneous rejections (false positives) that could trigger unnecessary system shutdowns, or, conversely, undetected novelties (false negatives) leading to silent, dangerous failures. The challenge of managing false positives, which can erode user trust and increase human workload, is addressed by human-in-the-loop frameworks that adaptively control the False Positive Rate (FPR) with theoretical guarantees (118). From a broader socio-technical perspective, the integration of human oversight in such systems also raises ethical questions about the cognitive load, potential for automation bias, and psychological impact on human supervisors, necessitating careful design of human-AI interfaces and clear protocols.

Beyond error rates, the very definition of "out-of-distribution" carries significant ethical weight. (29) critically argues that focusing solely on "Out-of-Distribution Detection" might be insufficient for safety, proposing "Out-of-Model-Scope" (OMS) detection as a more ethically aligned objective. OMS aims to identify inputs that would lead to actual model errors, rather than just distribution shifts, thereby directly addressing the imperative to abstain from unsafe predictions. Furthermore, the robustness of OOD detectors against malicious inputs is a critical safety concern, as adversarial attacks could manipulate detectors into making unsafe decisions (11). The inherent difficulty in distinguishing harmless from potentially unsafe OOD events, particularly in dynamic environments like Reinforcement Learning, underscores the need for clear definitions of "unknown events" and robust safety assurance frameworks for ML components (147). This highlights the necessity for ethical guidelines and potentially regulatory standards to govern the certification and deployment of OOD-enabled AI systems.

A particularly critical ethical dimension is the potential for bias in OOD detection, which can lead to unfair or discriminatory outcomes. If OOD models are trained on data reflecting societal biases, they can inadvertently amplify these biases. For instance, (8) demonstrates how spurious correlations in training data (e.g., associating certain backgrounds with specific classes) can severely degrade OOD detection performance. Models relying on these non-causal features might confidently misclassify inputs from underrep-

resented demographic groups as "anomalous" if those inputs exhibit features statistically correlated with OOD data in the training set. This can result in discriminatory rejections or differential treatment, where certain groups are disproportionately flagged as "outliers." Such biases are not merely technical failures but ethical breaches, demanding fairness-aware OOD algorithms that explicitly analyze and mitigate performance disparities across demographic subgroups. While interpretability methods like GAIA, which uses gradient-based attribution abnormality (77), or Neuron Activation Coverage (NAC) (30), are not direct fairness interventions, they are crucial tools for auditing models. By revealing *why* an input is deemed OOD, they enable practitioners to identify and address unintended biases in the OOD decision-making process. In multimodal contexts, where biases can exist across various data streams (e.g., text, video, audio), the challenge of ensuring fair OOD detection is further compounded (88).

Finally, the societal impact of deploying AI models that may fail silently on novel inputs is a pervasive ethical concern. The fundamental purpose of OOD detection is to prevent such silent failures, enabling models to express uncertainty or abstain when confronted with unfamiliar data. Methods like DoSE (12) directly tackle the "high likelihood for OOD" pathology, where generative models might assign high confidence to OOD data, thereby preventing a dangerous false sense of security. Furthermore, a deeper understanding of how in-distribution (ID) labels influence OOD detection, especially for "near OOD" scenarios where ethical risks are heightened (121), is vital to avoid mischaracterizing subtle shifts as benign. The development of robust and comprehensive in-distribution representations, as exemplified by methods like MOODv2 (115), inherently makes OOD detection more reliable and less prone to silent failures, as models gain a more accurate understanding of what constitutes "normal" data.

In conclusion, while significant technical advancements have propelled OOD detection forward, the ethical considerations and societal impact remain paramount. Future research must prioritize the development of robust safeguards, including statistical guarantees (as discussed in Section 7.2) and adaptive human-in-the-loop mechanisms (118), to rigorously control false positives and negatives in safety-critical applications. Crucially,

a concerted effort is needed to ensure fairness by investigating and mitigating potential biases, particularly those arising from spurious correlations (8), through the development of transparent and interpretable methods (77; 30) that facilitate auditing and accountability. Ultimately, the goal is to cultivate a paradigm where AI systems not only achieve high performance but also operate responsibly, recognizing their limitations, communicating uncertainty effectively, and adhering to ethical guidelines, thereby fostering trust and enabling the safe and equitable integration of AI into society.

References

References

[1] Te Han, and Yanfang Li (2022). *Out-of-distribution detection-assisted trustworthy machinery fault diagnosis approach with uncertainty-aware deep ensembles.* Reliability Engineering System Safety.

[2] Yibo Zhou (2022). *Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection.* Computer Vision and Pattern Recognition.

[3] Jingkang Yang, Kaiyang Zhou, and Ziwei Liu (2022). *Full-Spectrum Out-of-Distribution Detection.* International Journal of Computer Vision.

[4] E. Zisselman, and Aviv Tamar (2020). *Deep Residual Flow for Out of Distribution Detection.* Computer Vision and Pattern Recognition.

[5] Yue Song, N. Sebe, and Wei Wang (2022). *RankFeat: Rank-1 Feature Removal for Out-of-distribution Detection.* Neural Information Processing Systems.

[6] Yixin Liu, Kaize Ding, Huan Liu, et al. (2022). *GOOD-D: On Unsupervised Graph Out-Of-Distribution Detection.* Web Search and Data Mining.

[7] Alireza Zaeemzadeh, Niccoló Bisagno, Zeno Sambugaro, et al. (2021). *Out-of-Distribution Detection Using Union of 1-Dimensional Subspaces.* Computer Vision and Pattern Recognition.

[8] Yifei Ming, Hang Yin, and Yixuan Li (2021). *On the Impact of Spurious Correlation for Out-of-distribution Detection.* AAAI Conference on Artificial Intelligence.

[9] Yao Zhu, YueFeng Chen, Chuanlong Xie, et al. (2022). *Boosting Out-of-distribution Detection with Typical Features.* Neural Information Processing Systems.

[10] Taewon Jeong, and Heeyoung Kim (2020). *OOD-MAML: Meta-Learning for Few-Shot Out-of-Distribution Detection and Classification.* Neural Information Processing Systems.

[11] Jiefeng Chen, Yixuan Li, Xi Wu, et al. (2020). *Robust Out-of-distribution Detection for Neural Networks*. Unpublished manuscript.

[12] W. Morningstar, Cusuh Ham, Andrew Gallagher, et al. (2020). *Density of States Estimation for Out-of-Distribution Detection*. International Conference on Artificial Intelligence and Statistics.

[13] Zenan Li, Qitian Wu, Fan Nie, et al. (2022). *GraphDE: A Generative Framework for Debiased Learning and Out-of-Distribution Detection on Graphs*. Neural Information Processing Systems.

[14] W. Xie, Te Han, Zhong Pei, et al. (2023). *A unified out-of-distribution detection framework for trustworthy prognostics and health management in renewable energy systems*. Engineering applications of artificial intelligence.

[15] Yuyuan Liu, Choubo Ding, Yu Tian, et al. (2022). *Residual Pattern Learning for Pixel-wise Out-of-Distribution Detection in Semantic Segmentation*. IEEE International Conference on Computer Vision.

[16] David Zimmerer, Peter M. Full, Fabian Isensee, et al. (2022). *MOOD 2020: A Public Benchmark for Out-of-Distribution Detection and Localization on Medical Images*. IEEE Transactions on Medical Imaging.

[17] R. Kaur, Susmit Jha, Anirban Roy, et al. (2022). *iDECODe: In-distribution Equivariance for Conformal Out-of-distribution Detection*. AAAI Conference on Artificial Intelligence.

[18] Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, et al. (2021). *Mixture Outlier Exposure: Towards Out-of-Distribution Detection in Fine-grained Environments*. IEEE Workshop/Winter Conference on Applications of Computer Vision.

[19] Xin Dong, Junfeng Guo, Ang Li, et al. (2021). *Neural Mean Discrepancy for Efficient Out-of-Distribution Detection*. Computer Vision and Pattern Recognition.

[20] Feiyang Cai, and X. Koutsoukos (2020). *Real-time Out-of-distribution Detection in Learning-Enabled Cyber-Physical Systems.* International Conference on Cyber-Physical Systems.

[21] J. Gawlikowski, Sudipan Saha, Anna M. Kruspe, et al. (2022). *An Advanced Dirichlet Prior Network for Out-of-Distribution Detection in Remote Sensing.* IEEE Transactions on Geoscience and Remote Sensing.

[22] Unknown Authors (2020). *Hyperparameter-Free Out-of-Distribution Detection Using Cosine Similarity.* Asian Conference on Computer Vision.

[23] Konstantin Kirchheim, Marco Filax, and F. Ortmeier (2022). *PyTorch-OOD: A Library for Out-of-Distribution Detection based on PyTorch.* 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[24] Hao Lang, Yinhe Zheng, Yixuan Li, et al. (2023). *A Survey on Out-of-Distribution Detection in NLP.* Trans. Mach. Learn. Res..

[25] Yeonguk Yu, Sungho Shin, Seongju Lee, et al. (2022). *Block Selection Method for Using Feature Norm in Out-of-Distribution Detection.* Computer Vision and Pattern Recognition.

[26] Qingling Zhao, Mingqiang Chen, Zonghua Gu, et al. (2022). *CAN Bus Intrusion Detection Based on Auxiliary Classifier GAN and Out-of-distribution Detection.* ACM Transactions on Embedded Computing Systems.

[27] Mouin Ben Ammar, Nacim Belkhir, Sebastian Popescu, et al. (2023). *NECO: NEural Collapse Based Out-of-distribution detection.* International Conference on Learning Representations.

[28] Victor Besnier, Andrei Bursuc, David Picard, et al. (2021). *Triggering Failures: Out-Of-Distribution detection by learning from local adversarial attacks in Semantic Segmentation.* IEEE International Conference on Computer Vision.

[29] Joris Gu'erin, Kevin Delmas, Raul Sena Ferreira, et al. (2022). *Out-Of-Distribution Detection Is Not All You Need.* AAAI Conference on Artificial Intelligence.

[30] Y. Liu, Chris Xing Tian, Haoliang Li, et al. (2023). *Neuron Activation Coverage: Rethinking Out-of-distribution Detection and Generalization.* International Conference on Learning Representations.

[31] Christoph Berger, Magdalini Paschali, Ben Glocker, et al. (2021). *Confidence-based Out-of-Distribution Detection: A Comparative Study and Analysis.* UNSURE/PIPPI@MICCAI.

[32] Yijun Yang, Ruiyuan Gao, and Qiang Xu (2022). *Out-of-Distribution Detection with Semantic Mismatch under Masking.* European Conference on Computer Vision.

[33] Fan Lu, Kai Zhu, Wei Zhai, et al. (2023). *Uncertainty-Aware Optimal Transport for Semantically Coherent Out-of-Distribution Detection.* Computer Vision and Pattern Recognition.

[34] B. Kim, B. Kim, and Y. Hyun (2024). *Investigation of out-of-distribution detection across various models and training methodologies.* Neural Networks.

[35] Wenjun Miao, Guansong Pang, Tianqi Li, et al. (2023). *Out-of-Distribution Detection in Long-Tailed Recognition with Calibrated Outlier Class Learning.* AAAI Conference on Artificial Intelligence.

[36] Qizhou Wang, Feng Liu, Yonggang Zhang, et al. (2022). *Watermarking for Out-of-distribution Detection.* Neural Information Processing Systems.

[37] Ji Zhang, Lianli Gao, Bingguang Hao, et al. (2023). *From Global to Local: Multi-Scale Out-of-Distribution Detection.* IEEE Transactions on Image Processing.

[38] Jo-Lan Kuan, and Jonas W. Mueller (2022). *Back to the Basics: Revisiting Out-of-Distribution Detection Baselines.* arXiv.org.

[39] Hyunjun Choi, Hawook Jeong, and Jin Young Choi (2023). *Balanced Energy Regularization Loss for Out-of-distribution Detection.* Computer Vision and Pattern Recognition.

[40] Bojun Liu, Jordan G Boysen, I. C. Unarta, et al. (2025). *Exploring transition states of protein conformational changes via out-of-distribution detection in the hyperspherical latent space.* Nature Communications.

[41] Atsuyuki Miyai, Qing Yu, Go Irie, et al. (2023). *GL-MCM: Global and Local Maximum Concept Matching for Zero-Shot Out-of-Distribution Detection.* International Journal of Computer Vision.

[42] Julian Bitterwolf, Alexander Meinke, Maximilian Augustin, et al. (2022). *Breaking Down Out-of-Distribution Detection: Many Methods Based on OOD Training Data Estimate a Combination of the Same Core Quantities.* International Conference on Machine Learning.

[43] Eduardo Dadalto Camara Gomes, F. Alberge, P. Duhamel, et al. (2022). *Igeood: An Information Geometry Approach to Out-of-Distribution Detection.* International Conference on Learning Representations.

[44] Ruiyuan Gao, Chenchen Zhao, Lanqing Hong, et al. (2023). *DiffGuard: Semantic Mismatch-Guided Out-of-Distribution Detection using Pre-trained Diffusion Models.* IEEE International Conference on Computer Vision.

[45] Senqi Cao, and Zhongfei Zhang (2022). *Deep Hybrid Models for Out-of-Distribution Detection.* Computer Vision and Pattern Recognition.

[46] Tong Wei, Bo-Lin Wang, and Min-Ling Zhang (2023). *EAT: Towards Long-Tailed Out-of-Distribution Detection.* AAAI Conference on Artificial Intelligence.

[47] Sima Behpour, T. Doan, Xin Li, et al. (2023). *GradOrth: A Simple yet Efficient Out-of-Distribution Detection with Orthogonal Projection of Gradients.* Neural Information Processing Systems.

[48] Danny Wang, Ruihong Qiu, Guangdong Bai, et al. (2025). *GOLD: Graph Out-of-Distribution Detection via Implicit Adversarial Latent Generation*. International Conference on Learning Representations.

[49] M. Graham, W. H. Pinaya, P. Wright, et al. (2023). *Unsupervised 3D out-of-distribution detection with latent diffusion models*. International Conference on Medical Image Computing and Computer-Assisted Intervention.

[50] Tianyi Bao, Qitian Wu, Zetian Jiang, et al. (2024). *Graph Out-of-Distribution Detection Goes Neighborhood Shaping*. International Conference on Machine Learning.

[51] Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, et al. (2023). *Understanding the Feature Norm for Out-of-Distribution Detection*. IEEE International Conference on Computer Vision.

[52] Tom Haider, Karsten Roscher, Felippe Schmoeller da Roza, et al. (2023). *Out-of-Distribution Detection for Reinforcement Learning Agents with Probabilistic Dynamics Models*. Adaptive Agents and Multi-Agent Systems.

[53] Soumya Suvra Ghosal, Yiyou Sun, and Yixuan Li (2023). *How to Overcome Curse-of-Dimensionality for Out-of-Distribution Detection?*. AAAI Conference on Artificial Intelligence.

[54] William Yang, Byron Zhang, and Olga Russakovsky (2023). *ImageNet-OOD: Deciphering Modern Out-of-Distribution Detection Algorithms*. International Conference on Learning Representations.

[55] Sishuo Chen, Wenkai Yang, Xiaohan Bi, et al. (2023). *Fine-Tuning Deteriorates General Textual Out-of-Distribution Detection by Distorting Task-Agnostic Features*. Findings.

[56] Reza Averly, and Wei-Lun Chao (2023). *Unified Out-Of-Distribution Detection: A Model-Specific Perspective*. IEEE International Conference on Computer Vision.

[57] Jianing Zhu, Hengzhuang Li, Jiangchao Yao, et al. (2023). *Unleashing Mask: Explore the Intrinsic Out-of-Distribution Detection Capability.* International Conference on Machine Learning.

[58] Divyanshu Mishra, He Zhao, Pramit Saha, et al. (2023). *Dual Conditioned Diffusion Models for Out-of-Distribution Detection: Application to Fetal Ultrasound Videos.* International Conference on Medical Image Computing and Computer-Assisted Intervention.

[59] Shuyang Yu, Junyuan Hong, Haotao Wang, et al. (2023). *Turning the Curse of Heterogeneity in Federated Learning into a Blessing for Out-of-Distribution Detection.* International Conference on Learning Representations.

[60] Yi Dai, Hao Lang, Kaisheng Zeng, et al. (2023). *Exploring Large Language Models for Multi-Modal Out-of-Distribution Detection.* Conference on Empirical Methods in Natural Language Processing.

[61] Teresa Araújo, Guilherme Aresta, U. Schmidt-Erfurth, et al. (2023). *Few-shot out-of-distribution detection for automated screening in retinal OCT images using deep learning.* Scientific Reports.

[62] Ming Xu, Zheng Lian, B. Liu, et al. (2023). *VRA: Variational Rectified Activation for Out-of-distribution Detection.* Neural Information Processing Systems.

[63] Zhen Cheng, Fei Zhu, Xu-Yao Zhang, et al. (2023). *Average of Pruning: Improving Performance and Stability of Out-of-Distribution Detection.* IEEE Transactions on Neural Networks and Learning Systems.

[64] Anton Vasiliuk, Daria Frolova, M. Belyaev, et al. (2023). *Limitations of Out-of-Distribution Detection in 3D Medical Image Segmentation.* Journal of Imaging.

[65] Harry Anthony, and K. Kamnitsas (2023). *On the use of Mahalanobis distance for out-of-distribution detection with neural networks for medical imaging.* UNSURE@MICCAI.

[66] Taocun Yang, Y. Huang, Yanlin Xie, et al. (2023). *MixOOD: Improving Out-of-distribution Detection with Enhanced Data Mixup*. ACM Trans. Multim. Comput. Commun. Appl..

[67] Bo Liu, Li-Ming Zhan, Zexin Lu, et al. (2023). *How Good Are LLMs at Out-of-Distribution Detection?*. International Conference on Language Resources and Evaluation.

[68] Xiaoyuan Guan, Zhouwu Liu, Weishi Zheng, et al. (2023). *Revisit PCA-based technique for Out-of-Distribution Detection*. IEEE International Conference on Computer Vision.

[69] Zihan Zhang, Zhuo Xu, and Xiang Xiang (2024). *Vision-Language Dual-Pattern Matching for Out-of-Distribution Detection*. European Conference on Computer Vision.

[70] Kai Liu, Zhihang Fu, Chao Chen, et al. (2024). *Category-Extensible Out-of-Distribution Detection via Hierarchical Context Descriptions*. Neural Information Processing Systems.

[71] Yulong Jia, Jiaming Li, Ganlong Zhao, et al. (2024). *Enhancing out-of-distribution detection via diversified multi-prototype contrastive learning*. Pattern Recognition.

[72] Wenyu Jiang, Hao Cheng, Mingcai Chen, et al. (2023). *DOS: Diverse Outlier Sampling for Out-of-Distribution Detection*. International Conference on Learning Representations.

[73] Zhitong Gao, Shipeng Yan, and Xuming He (2023). *ATTA: Anomaly-aware Test-Time Adaptation for Out-of-Distribution Detection in Segmentation*. Neural Information Processing Systems.

[74] Jens Henriksson, Stig Ursing, Murat Erdogan, et al. (2023). *Out-of-Distribution Detection as Support for Autonomous Driving Safety Lifecycle*. Requirements Engineering: Foundation for Software Quality.

[75] M. Saadati, Aditya Balu, Shivani Chiranjeevi, et al. (2023). *Out-of-Distribution Detection Algorithms for Robust Insect Classification*. Plant Phenomics.

[76] Gongxun Miao, Guohua Wu, Zhen Zhang, et al. (2023). *SPN: A Method of Few-Shot Traffic Classification With Out-of-Distribution Detection Based on Siamese Prototypical Network*. IEEE Access.

[77] Jinggang Chen, Junjie Li, Xiaoyang Qu, et al. (2023). *GAIA: Delving into Gradient-based Attribution Abnormality for Out-of-distribution Detection*. Neural Information Processing Systems.

[78] Eduardo Aguilar, B. Raducanu, P. Radeva, et al. (2023). *Continual Evidential Deep Learning for Out-of-Distribution Detection*. 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).

[79] Yawen Ouyang, Yongchang Cao, Yuan Gao, et al. (2023). *On Prefix-tuning for Lightweight Out-of-distribution Detection*. Annual Meeting of the Association for Computational Linguistics.

[80] Marc Lafon, Elias Ramzi, Clément Rambour, et al. (2023). *Hybrid Energy Based Model in the Feature Space for Out-of-Distribution Detection*. International Conference on Machine Learning.

[81] Michael Kösel, M. Schreiber, Michael Ulrich, et al. (2024). *Revisiting Out-of-Distribution Detection in LiDAR-based 3D Object Detection*. 2024 IEEE Intelligent Vehicles Symposium (IV).

[82] Choubo Ding, and Guansong Pang (2024). *Zero-Shot Out-of-Distribution Detection with Outlier Label Exposure*. IEEE International Joint Conference on Neural Network.

[83] Yonggang Zhang, Jie Lu, Bo Peng, et al. (2024). *Learning to Shape In-distribution Feature Space for Out-of-distribution Detection*. Neural Information Processing Systems.

[84] Yingwen Wu, Ruiji Yu, Xinwen Cheng, et al. (2024). *Pursuing Feature Separation based on Neural Collapse for Out-of-Distribution Detection*. International Conference on Learning Representations.

[85] Haiyu Yao, Zongbo Han, Huazhu Fu, et al. (2024). *Out-Of-Distribution Detection with Diversification (Provably)*. Neural Information Processing Systems.

[86] Chao Chen, Zhihang Fu, Kai Liu, et al. (2024). *Optimal Parameter and Neuron Pruning for Out-of-Distribution Detection*. Neural Information Processing Systems.

[87] Ramneet Kaur, Yahan Yang, O. Sokolsky, et al. (2024). *Out-of-distribution Detection in Dependent Data for Cyber-physical Systems with Conformal Guarantees*. ACM Trans. Cyber Phys. Syst..

[88] Hanlei Zhang, Qianrui Zhou, Hua Xu, et al. (2024). *Multimodal Classification and Out-of-distribution Detection for Multimodal Intent Understanding*. arXiv.org.

[89] Sebastian Schmidt, Leonard Schenk, Leo Schwinn, et al. (2024). *A Unified Approach Towards Active Learning and Out-of-Distribution Detection*. Trans. Mach. Learn. Res..

[90] Jun Nie, Yadan Luo, Shanshan Ye, et al. (2024). *Out-of-Distribution Detection with Virtual Outlier Smoothing*. International Journal of Computer Vision.

[91] Haoyan Xu, Zhengtao Yao, Yushun Dong, et al. (2025). *Few-Shot Graph Out-of-Distribution Detection with LLMs*. arXiv.org.

[92] Tomás Vojír, Jan Sochman, Rahaf Aljundi, et al. (2023). *Calibrated Out-of-Distribution Detection with a Generic Representation*. 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW).

[93] Haodong Lu, Dong Gong, Shuo Wang, et al. (2024). *Learning with Mixture of Prototypes for Out-of-Distribution Detection*. International Conference on Learning Representations.

[94] Jun Nie, Yonggang Zhang, Zhen Fang, et al. (2024). *Out-of-Distribution Detection with Negative Prompts*. International Conference on Learning Representations.

[95] Xuefeng Du, Zhen Fang, Ilias Diakonikolas, et al. (2024). *How Does Unlabeled Data Provably Help Out-of-Distribution Detection?*. International Conference on Learning Representations.

[96] Tianqi Li, Guansong Pang, Xiaolong Bai, et al. (2024). *Learning Transferable Negative Prompts for Out-of-Distribution Detection*. Computer Vision and Pattern Recognition.

[97] Xiang Fang, A. Easwaran, B. Genest, et al. (2024). *Your data is not perfect: Towards cross-domain out-of-distribution detection in class-imbalanced data*. Expert systems with applications.

[98] J. Linmans, Gabriel Raya, J. Laak, et al. (2024). *Diffusion models for out-of-distribution detection in digital pathology*. Medical Image Anal..

[99] Jiaqi Chen, T. H. Teo, C. Kok, et al. (2024). *A Novel Single-Word Speech Recognition on Embedded Systems Using a Convolution Neuron Network with Improved Out-of-Distribution Detection*. Electronics.

[100] Hao Dong, Yue Zhao, Eleni Chatzi, et al. (2024). *MultiOOD: Scaling Out-of-Distribution Detection for Multiple Modalities*. Neural Information Processing Systems.

[101] Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, et al. (2024). *Generalized Out-of-Distribution Detection and Beyond in Vision Language Model Era: A Survey*. Trans. Mach. Learn. Res..

[102] Xiaochen Zhang, Chen Wang, Wei Zhou, et al. (2024). *Trustworthy Diagnostics With Out-of-Distribution Detection: A Novel Max-Consistency and Min-Similarity Guided Deep Ensembles for Uncertainty Estimation*. IEEE Internet of Things Journal.

[103] Chentao Cao, Zhun Zhong, Zhanke Zhou, et al. (2024). *Envisioning Outlier Exposure by Large Language Models for Out-of-Distribution Detection*. International Conference on Machine Learning.

[104] Bo Peng, Yadan Luo, Yonggang Zhang, et al. (2024). *ConjNorm: Tractable Density Estimation for Out-of-Distribution Detection*. International Conference on Learning Representations.

[105] Luzhi Wang, Dongxiao He, He Zhang, et al. (2024). *GOODAT: Towards Test-time Graph Out-of-Distribution Detection*. AAAI Conference on Artificial Intelligence.

[106] Yili Wang, Yixin Liu, Xu Shen, et al. (2024). *Unifying Unsupervised Graph-Level Anomaly Detection and Out-of-Distribution Detection: A Benchmark*. International Conference on Learning Representations.

[107] Yiming Wang, Pei Zhang, Baosong Yang, et al. (2024). *Embedding Trajectory for Out-of-Distribution Detection in Mathematical Reasoning*. Neural Information Processing Systems.

[108] Zesheng Hong, Yubiao Yue, Yubin Chen, et al. (2024). *Out-of-distribution Detection in Medical Image Analysis: A survey*. arXiv.org.

[109] Geng Yu, Jianing Zhu, Jiangchao Yao, et al. (2024). *Self-Calibrated Tuning of Vision-Language Models for Out-of-Distribution Detection*. Neural Information Processing Systems.

[110] Ke Fan, Tong Liu, Xingyu Qiu, et al. (2024). *Test-Time Linear Out-of-Distribution Detection*. Computer Vision and Pattern Recognition.

[111] Kun Fang, Qinghua Tao, Kexin Lv, et al. (2024). *Kernel PCA for Out-of-Distribution Detection*. Neural Information Processing Systems.

[112] Tom'avs Voj'ivr, Jan Sochman, and Jivr'i Matas (2024). *PixOOD: Pixel-Level Out-of-Distribution Detection*. European Conference on Computer Vision.

[113] Li Li, Huixian Gong, Hao Dong, et al. (2024). *DPU: Dynamic Prototype Updating for Multimodal Out-of-Distribution Detection*. arXiv.org.

[114] Keke Tang, Chao Hou, Weilong Peng, et al. (2024). *CORES: Convolutional Response-based Score for Out-of-distribution Detection*. Computer Vision and Pattern Recognition.

[115] Jingyao Li, Pengguang Chen, Shaozuo Yu, et al. (2024). *MOODv2: Masked Image Modeling for Out-of-Distribution Detection*. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[116] Claus Hofmann, Simon Schmid, Bernhard Lehner, et al. (2024). *Energy-based Hopfield Boosting for Out-of-Distribution Detection*. Neural Information Processing Systems.

[117] Zhi Zhou, Ming Yang, Jiang-Xin Shi, et al. (2024). *DeCoOp: Robust Prompt Tuning with Out-of-Distribution Detection*. International Conference on Machine Learning.

[118] Harit Vishwakarma, Heguang Lin, and Ramya Korlakai Vinayak (2024). *Taming False Positives in Out-of-Distribution Detection with Human Feedback*. International Conference on Artificial Intelligence and Statistics.

[119] Kaizheng Wang, Fabio Cuzzolin, Keivan K1 Shariatmadar, et al. (2024). *Credal Wrapper of Model Averaging for Uncertainty Estimation on Out-Of-Distribution Detection*. arXiv.org.

[120] Ruiyao Xu, and Kaize Ding (2024). *Large Language Models for Anomaly and Out-of-Distribution Detection: A Survey*. North American Chapter of the Association for Computational Linguistics.

[121] Xuefeng Du, Yiyou Sun, and Yixuan Li (2024). *When and How Does In-Distribution Label Help Out-of-Distribution Detection?*. International Conference on Machine Learning.

[122] Ghada Zamzmi, Kesavan Venkatesh, Brandon Nelson, et al. (2024). *Out-of-Distribution Detection and Radiological Data Monitoring Using Statistical Process Control*. Journal of imaging informatics in medicine.

[123] Armando Zhu, Jiabei Liu, Keqin Li, et al. (2024). *Exploiting Diffusion Prior for Out-of-Distribution Detection*. Irish Interdisciplinary Journal of Science amp; Research.

[124] L. Nasvytis, Kai Sandbrink, Jakob Foerster, et al. (2024). *Rethinking Out-of-Distribution Detection for Reinforcement Learning: Advancing Methods for Evaluation and Detection*. Adaptive Agents and Multi-Agent Systems.

[125] Yue Yuan, Rundong He, Yicong Dong, et al. (2024). *Discriminability-Driven Channel Selection for Out-of-Distribution Detection*. Computer Vision and Pattern Recognition.

[126] Zhen Fang, Yixuan Li, Feng Liu, et al. (2024). *On the Learnability of Out-of-distribution Detection*. Journal of machine learning research.

[127] Yanan Cao, Fengzhao Shi, Qing Yu, et al. (2025). *IBPL: Information Bottleneck-based Prompt Learning for graph out-of-distribution detection*. Neural Networks.

[128] Alvin Heng, A. Thiéry, and Harold Soh (2024). *Out-of-Distribution Detection with a Single Unconditional Diffusion Model*. Neural Information Processing Systems.

[129] Qinyu Zhao, Ming Xu, Kartik Gupta, et al. (2024). *Towards Optimal Feature-Shaping Methods for Out-of-Distribution Detection*. International Conference on Learning Representations.

[130] Chenhui Xu, Fuxun Yu, Zirui Xu, et al. (2024). *Out-of-Distribution Detection via Deep Multi-Comprehension Ensemble*. International Conference on Machine Learning.

[131] Hossein Mirzaei, and Mackenzie W. Mathis (2024). *Adversarially Robust Out-of-Distribution Detection Using Lyapunov-Stabilized Embeddings*. International Conference on Learning Representations.

[132] Sina Sharifi, Taha Entesari, Bardia Safaei, et al. (2024). *Gradient-Regularized Out-of-Distribution Detection.* European Conference on Computer Vision.

[133] Konstantin Kirchheim, Tim Gonschorek, and F. Ortmeier (2024). *Out-of-Distribution Detection with Logical Reasoning.* IEEE Workshop/Winter Conference on Applications of Computer Vision.

[134] Xiangxi Shi, and Stefan Lee (2024). *Benchmarking Out-of-Distribution Detection in Visual Question Answering.* IEEE Workshop/Winter Conference on Applications of Computer Vision.

[135] Shuai Feng, and Chongjun Wang (2024). *When an extra rejection class meets out-of-distribution detection in long-tailed image classification.* Neural Networks.

[136] Yixia Li, Boya Xiong, Guanhua Chen, et al. (2024). *SeTAR: Out-of-Distribution Detection with Selective Low-Rank Approximation.* Neural Information Processing Systems.

[137] Fanhu Zeng, Zhen Cheng, Fei Zhu, et al. (2024). *Local-Prompt: Extensible Local Prompts for Few-Shot Out-of-Distribution Detection.* International Conference on Learning Representations.

[138] Gerhard Krumpl, H. Avenhaus, Horst Possegger, et al. (2024). *ATS: Adaptive Temperature Scaling for Enhancing Out-of-Distribution Detection Methods.* IEEE Workshop/Winter Conference on Applications of Computer Vision.

[139] L. Hogeweg, Rajesh Gangireddy, Django Brunink, et al. (2024). *COOD: Combined out-of-distribution detection using multiple measures for anomaly novel class detection in large-scale hierarchical classification.* 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[140] Hongjun Wang, S. Vaze, and Kai Han (2024). *Dissecting Out-of-Distribution Detection and Open-Set Recognition: A Critical Analysis of Methods and Benchmarks.* International Journal of Computer Vision.

[141] Shuo Lu, Yingsheng Wang, Lijun Sheng, et al. (2024). *Out-of-Distribution Detection: A Task-Oriented Survey of Recent Advances*. Unpublished manuscript.

[142] Qiaozhi Tan, Long Bai, Guan-Feng Wang, et al. (2024). *Endoood: Uncertainty-Aware Out-of-Distribution Detection in Capsule Endoscopy Diagnosis*. IEEE International Symposium on Biomedical Imaging.

[143] Longfei Ma, Yiyou Sun, Kaize Ding, et al. (2024). *Revisiting Score Propagation in Graph Out-of-Distribution Detection*. Neural Information Processing Systems.

[144] Shenzhi Yang, Bin Liang, An Liu, et al. (2025). *Bounded and Uniform Energy-based Out-of-distribution Detection for Graphs*. International Conference on Machine Learning.

[145] Yuhang Zhang, Jiani Hu, Dongchao Wen, et al. (2024). *Unsupervised evaluation for out-of-distribution detection*. Pattern Recognition.

[146] Lemar Abdi, M. Valiuddin, Christiaan G. A. Viviers, et al. (2024). *Typicality Excels Likelihood for Unsupervised Out-of-Distribution Detection in Medical Imaging*. UNSURE@MICCAI.

[147] Tom Haider, Karsten Roscher, Benjamin Herd, et al. (2024). *Can you trust your Agent? The Effect of Out-of-Distribution Detection on the Safety of Reinforcement Learning Systems*. ACM Symposium on Applied Computing.

[148] Jingen Qu, Yufei Chen, Xiaodong Yue, et al. (2024). *Hyper-opinion Evidential Deep Learning for Out-of-Distribution Detection*. Neural Information Processing Systems.

[149] Paul Novello, Joseba Dalmau, and L'eo Andeol (2024). *Out-of-Distribution Detection Should Use Conformal Prediction (and Vice-versa?)*. arXiv.org.

[150] Jiankang Chen, Tong Zhang, Weishi Zheng, et al. (2024). *TagFog: Textual Anchor Guidance and Fake Outlier Generation for Visual Out-of-Distribution Detection*. AAAI Conference on Artificial Intelligence.

[151] Wenjun Miao, Guansong Pang, Jingyi Zheng, et al. (2024). *Long-Tailed Out-of-Distribution Detection via Normalized Outlier Distribution Adaptation.* Neural Information Processing Systems.

[152] Ji-Hun Oh, Kianoush Falahkheirkhah, and Rohit Bhargava (2024). *Are We Ready for Out-of-Distribution Detection in Digital Pathology?.* International Conference on Medical Image Computing and Computer-Assisted Intervention.

[153] Jiuqing Dong, Yifan Yao, Wei Jin, et al. (2024). *Enhancing Few-Shot Out-of-Distribution Detection With Pre-Trained Model Features.* IEEE Transactions on Image Processing.

[154] Shuai Feng, Pengsheng Jin, and Chongjun Wang (2024). *CASE: Exploiting Intra-class Compactness and Inter-class Separability of Feature Embeddings for Out-of-Distribution Detection.* AAAI Conference on Artificial Intelligence.

[155] Rundong He, Yue Yuan, Zhongyi Han, et al. (2024). *Exploring Channel-Aware Typical Features for Out-of-Distribution Detection.* AAAI Conference on Artificial Intelligence.

[156] Mingrong Gong, Chaoqi Chen, Qingqiang Sun, et al. (2024). *Out-of-Distribution Detection with Prototypical Outlier Proxy.* AAAI Conference on Artificial Intelligence.

[157] Evan D. Cook, Marc-Antoine Lavoie, and Steven L. Waslander (2024). *Feature Density Estimation for Out-of-Distribution Detection via Normalizing Flows.* Proceedings of the 21st Conference on Robots and Vision.

[158] Wenjun Miao, Guansong Pang, Trong-Tung Nguyen, et al. (2024). *OpenCIL: Benchmarking Out-of-Distribution Detection in Class-Incremental Learning.* Pattern Recognition.

[159] Yuxiao Lee, Xiaofeng Cao, Jingcai Guo, et al. (2025). *Concept Matching with Agent for Out-of-Distribution Detection.* AAAI Conference on Artificial Intelligence.

[160] Jinglong Wang, and Ridong Zhang (2025). *Open-Set Fault Diagnosis Based on 1D-ResNet With Fusion of Cross-Class and Extreme Information for Out-of-Distribution Detection.* IEEE Transactions on Instrumentation and Measurement.

[161] Rundong He, Zhongyi Han, Xiushan Nie, et al. (2024). *Visual Out-of-Distribution Detection in Open-Set Noisy Environments.* International Journal of Computer Vision.

[162] Xuhui Li, Zhen Fang, Yonggang Zhang, et al. (2025). *Characterizing Submanifold Region for Out-of-Distribution Detection.* IEEE Transactions on Knowledge and Data Engineering.

[163] Aryan Gulati, Xingjian Dong, Carlos Hurtado, et al. (2024). *Out-of-Distribution Detection through Soft Clustering with Non-Negative Kernel Regression.* Conference on Empirical Methods in Natural Language Processing.

[164] Genki Osada, Tsubasa Takahashi, and Takashi Nishide (2024). *Understanding Likelihood of Normalizing Flow and Image Complexity through the Lens of Out-of-Distribution Detection.* AAAI Conference on Artificial Intelligence.

[165] Yihan Mei, Xinyu Wang, De-Fu Zhang, et al. (2024). *Multi-Label Out-of-Distribution Detection with Spectral Normalized Joint Energy.* APWeb/WAIM.

[166] Jeonghyeon Kim, Jihyo Kim, and Sangheum Hwang (2024). *Comparison of Out-of-Distribution Detection Performance of CLIP-based Fine-Tuning Methods.* International Conference on Electronics, Information and Communications.

[167] Sabri Mustafa Kahya, Boran Hamdi Sivrikaya, Muhammet Sami Yavuz, et al. (2024). *FOOD: Facial Authentication and Out-of-Distribution Detection with Short-Range FMCW Radar.* International Conference on Information Photonics.

[168] Burak Ekim, G. Tadesse, Caleb Robinson, et al. (2024). *Distribution Shifts at Scale: Out-of-distribution Detection in Earth Observation.* 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).

[169] Dong Geun Shin, and Hye Won Chung (2024). *Representation Norm Amplification for Out-of-Distribution Detection in Long-Tail Learning.* Trans. Mach. Learn. Res..

[170] Renmingyue Du, Jixun Yao, Qiuqiang Kong, et al. (2024). *Towards Out-of-Distribution Detection in Vocoder Recognition via Latent Feature Reconstruction.* Unpublished manuscript.

[171] Zhenjiang Mao, Dong-You Jhong, Ao Wang, et al. (2024). *Language-Enhanced Latent Representations for Out-of-Distribution Detection in Autonomous Driving.* arXiv.org.

[172] Francesco Cappio Borlino, L. Lu, and Tatiana Tommasi (2024). *Foundation Models and Fine-Tuning: A Benchmark for Out of Distribution Detection.* IEEE Access.

[173] SiCong Li, Ning Li, Min Jing, et al. (2024). *Evaluation of Ten Deep-Learning-Based Out-of-Distribution Detection Methods for Remote Sensing Image Scene Classification.* Remote Sensing.

[174] Qichao Chen, Kuan Li, Zhiyuan Chen, et al. (2024). *Exploring feature sparsity for out-of-distribution detection.* Scientific Reports.

[175] Jingqiu Zhou, Aojun Zhou, and Hongsheng Li (2024). *NODI: Out-Of-Distribution Detection with Noise from Diffusion.* arXiv.org.

[176] Silvio Galesso, Philipp Schröppel, Hssan Driss, et al. (2024). *Diffusion for Out-of-Distribution Detection on Road Scenes and Beyond.* European Conference on Computer Vision.

[177] Yang Chen, Chih-Li Sung, A. Kusari, et al. (2024). *Uncertainty-Aware Out-of-Distribution Detection with Gaussian Processes.* arXiv.org.

[178] Xingming Long, Jie Zhang, Shiguang Shan, et al. (2024). *Rethinking the Evaluation of Out-of-Distribution Detection: A Sorites Paradox.* Neural Information Processing Systems.

[179] Tingyi Cai, Yunliang Jiang, Yixin Liu, et al. (2025). *Out-of-Distribution Detection on Graphs: A Survey*. arXiv.org.

[180] Moru Liu, Hao Dong, Jessica Kelly, et al. (2025). *Extremely Simple Multimodal Outlier Synthesis for Out-of-Distribution Detection and Segmentation*. arXiv.org.

[181] Hengzhuang Li, and Teng Zhang (2025). *Outlier Synthesis via Hamiltonian Monte Carlo for Out-of-Distribution Detection*. International Conference on Learning Representations.

[182] Xixi Liu, and Christopher Zach (2024). *TAG: Text Prompt Augmentation for Zero-Shot Out-of-Distribution Detection*. European Conference on Computer Vision.

[183] Yeonguk Yu, Sungho Shin, Minhwan Ko, et al. (2024). *Exploring using jigsaw puzzles for out-of-distribution detection*. Computer Vision and Image Understanding.

[184] Syed Safwan Ahsan, Alireza Esmaeilzehi, and M. O. Ahmad (2024). *OODNet: A deep blind JPEG image compression deblocking network using out-of-distribution detection*. Journal of Visual Communication and Image Representation.

[185] Xinsong Ma, Xin Zou, and Weiwei Liu (2024). *A Provable Decision Rule for Out-of-Distribution Detection*. International Conference on Machine Learning.

[186] Erblin Isaku, Hassan Sartaj, and Shaukat Ali (2025). *Digital Twin-based Out-of-Distribution Detection in Autonomous Vessels*. arXiv.org.