Handling Out-of-Distribution Data: A Survey

Lakpa Tamang, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal

Abstract—In the field of Machine Learning (ML) and data-driven applications, one of the significant challenge is the change in data distribution between the training and deployment stages, commonly known as distribution shift. This paper outlines different mechanisms for handling two main types of distribution shifts: (i) **Covariate shift:** where the value of features or covariates change between train and test data, and (ii) **Concept/Semantic-shift:** where model experiences shift in the concept learned during training due to emergence of novel classes in the test phase. We sum up our contributions in three folds. First, we formalize distribution shifts, recite on how the conventional method fails to handle them adequately and urge for a model that can simultaneously perform better in all types of distribution shifts. Second, we discuss why handling distribution shifts is important and provide an extensive review of the methods and techniques that have been developed to detect, measure, and mitigate the effects of these shifts. Third, we discuss the current state of distribution shift handling mechanisms and propose future research directions in this area. Overall, we provide a retrospective synopsis of the literature in the distribution shift, focusing on OOD data that had been overlooked in the existing surveys.

Index Terms—Data Distribution Shift, Out-of-Distribution, Covariate Shift, Concept Shift

1 Introduction

Existing Machine Learning (ML) techniques, particularly Deep Neural Networks (DNNs), have shown unprecedented success across domains, such as computer vision, natural language processing, and recommendation systems [1]. These models tend to exploit subtle statistical correlations present in the training distribution, yielding impressive results under the *i.i.d* (independently and identically distributed) hypothesis. However, despite their prowess under controlled experimental conditions, there is growing empirical evidence highlighting their vulnerabilities to real-world data distribution shifts. These shifts may often surface in relation to several confining factors, such as sample selection biases, non-stationary environments, and other inherent peculiarities of data generation mechanisms [2]. As demonstrated by the adversarial examples in [3], even subtle changes in the data distribution can have a significant impact on the performance of advanced classifiers. Therefore, it is imperative to understand and address these vulnerabilities, especially for systems that perform safety-critical or high-impact operations such as medical diagnosis and autonomous vehicles.

The change in data distribution can hamper the model's accuracy, making its result unreliable to adapt to. One of several factors instigating these changes is the bias introduced by the experimental design due to the inimitability of the inconsistent testing conditions during training [4]. In other words, the knowledge that the model has learned from the training data may be sampled under conditions different from the ones it will encounter during actual testing. For graphical illustration, we refer to Fig. 1 where the large corpus of similarly and distinctly distributed test data (in feature, i.e., *covariates* space or semantic space) are put against a limited number of instances from the training sample space. While it is impractical to address all real-world test cases, it is crucial that the model be capable of handling such variations without compromising the relevance of the deployed model. Moreover, to make informed decisions about when and how to update the model, complying to the changes in the distribution of data that affect the model's performance is very important.

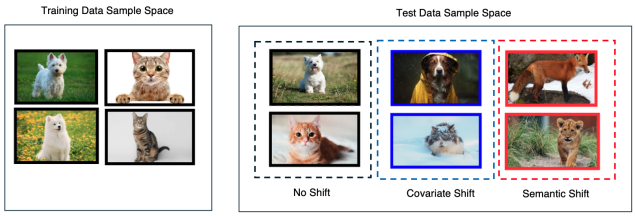


Fig. 1: Schematic illustration of data distribution shift between training and testing data sample spaces. In general, a model is trained on a limited knowledge of real-world samples, but it can encounter a whole set of similar or differently distributed inputs either in feature or semantic spaces when subjected to testing in the wild.

It is worthy to note that best practices for detecting shifts in high-dimensional real-world data have not yet been definitively established [5]. Regardless, numerous studies have been proposed with the primary objective of addressing the changes in data distribution by adapting and generalising to distributionally shifted samples or rejecting them entirely. In practice, the data distribution can be shifted in one of two ways: in feature space (*covariate shift*) or in label space (*semantic* or *concept shift*). Although numerous review papers have been released discussing the strategies for effectively addressing individual shifts, our review is the inaugural one to recognize these shifts as a collective issue. Rather than constraining the work into specifics of a single type of distributional shift [6], [7], we aim to emphasize the topic into a broader spectrum of research focusing on how each type of distributional shift is handled to make

School of Information Technology, Deakin University, Geelong Waurn Ponds Campus, Australia. E-mail: l.tamang@research.deakin.edu.au, {reda.bouadjenek, richard.dazeley, sunil.aryal}@deakin.edu.au
Manuscript received Xxx XX, 2025; revised Xxxx XX, 2025.

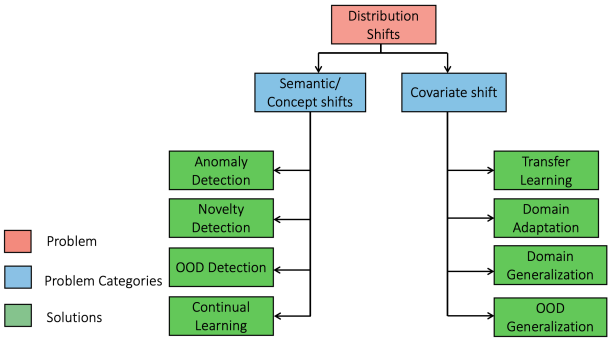


Fig. 2: Taxonomy of different methodologies for handling data distribution shift.

a model reliable in practice. In this paper, we present a comprehensive survey of the different modelling strategies dealing with data distribution change. Particularly, we aim to provide a comprehensive and nuanced understanding of the topic to the researchers by focusing on the retrospective overview of different methodologies centralized around handling data distribution shift problems as a generalized entity. In summary, through this paper, we offer three distinct contributions to the academic community in this domain of research:

1) **Formalization of Shifts** We formalize the prevalent types of shift and advocate that modelling strategies should account for both generalization and rejection when subjected to data distribution shifts. In accordance with this theory, we link existing topics by examining them independently to help the research community understand the practical objective of building ML models that are robust enough to handle both types of distribution shifts.
2) **Comprehensive Review of Modelling Strategies** In light of the presence of other literature reviews in the same field [6], [7], [8], [9] our work stands out by providing a comprehensive overview of key discoveries in the field of related topics, specifically centred around shift handling objective.
3) **Future Research Directions** We aim to provide readers with a deeper understanding of the current challenges and opportunities in this topic while also highlighting potential future research directions.

2 Background

2.1 Data Distribution

Data distribution refers to the arrangement of data values in a dataset and provides insights into patterns, characteristics, and relationships within the data [10]. It is crucial to understand data distribution that spans over a wide range of topics, including statistics, machine learning, and data analysis, in order to establish one of many objectives such as discovering trends, identifying outliers, and making informed decisions. The distribution of data obtained from a sample is vital in understanding how to analyze it, as it provides a parameterized mathematical function that can be used to calculate the probability for any individual observation from the sample space. In the field of statistics, different kinds of data distributions exist, such as normal [11], uniform [12], skewed [13], and bimodal [14], where each has their unique characteristics associated with the sample. In Machine learning and Deep Learning, probability distributions are considered to model real-world data and make predictions [15], [16]. A probability distribution [17] is a mathematical function that describes the likelihood of different outcomes for a random variable. It allows for the quantification of uncertainty and the making of predictions based on past data. These algorithms often involve estimating probability distributions from sample data and using them to generalize to new examples.

2.2 Why Data Distribution Changes?

There is an assumption that the distributions specified by unconditional or conditional models are static, remaining unchanged from the time they are learned to the time they are used [4]. However, if this assumption is not true and the distributions undergo some kind of change, then we must account for this change or at least the possibility of it. This requires examining the reasons why such a shift may occur. There are several reasons why an ML model might exhibit a data distribution shift.

Bias During Sample Selection: The concept of sample selection bias refers to a fault in the process of collecting or labelling data that leads to the uneven distribution of training examples. This results from the fact that the training examples were obtained through a biased method, which means they may not accurately reflect the environment where the classifier will be used. During the sampling process of the training data, the data points x_i^{te} may not precisely represent the actual testing distribution $P^{te}(X, Y)$. For instance, while generating a handwritten dataset, one may get rid of an entirely obscure character, although it may hold true that some characters are more likely to be written in an unclear manner.

Deployment Environment Changes: It is often true that data remains non-stationary to time and space change [18]. Environments are dynamic in general, and sometimes the difficulties of matching the learning scenario (with training data) to the real-world use (test data) are constrained by these changes. Such scenarios make it challenging to develop an understanding of the appropriateness of a particular model in the circumstance of these environmental changes, thus the prevalence of the shift. For example, a commonly observed issue with ML models trained to predict electricity demand based on historical data of usage time, temperature, and humidity is their failure when non-stationary changes such as climate and the adoption of renewable energy sources occur.

Change in the Domain: Occasionally, a new sample might be collected from a different domain to represent the same category. In this regard, various domains may use various terms to refer to the same entity. The changes in the domain are characterized by the fact that the measurement system or the description technique

Table 1: List of abbreviations used throughout the paper.

Abbreviations	Full forms
FID	Frechet Inception Distance
LPIPS	Learned Perceptual Image Patch Similarity
RMSE	Root Mean Squared Error
MAE	Mean Absolute Error
TPR	True Positive Rate
FPR	False Positive Rate
TNR	True Negative Rate
KID	Kernel Inception Distance
IoU	Intersection of Union
FDR	False Detection Rate
AUC	Area Under Curve
NMI	Normalized Mutual Index
AP	Average Precision
AUROC	Area Under Receiver Operating Characteristic Curve
AUPR	Area Under Precision Recall Curve
FPR	False Positive Rate
OSCR	Open Set Classification Rate
BWF	Backward Forgetting
FWT	Forward Transfer
BWT	Backward Transfer
MCR	Mean Class Recall

of the feature of a dataset is changed. For example, in a computer vision scenario, the change in visual concepts such as illumination conditions, image resolution or background of x_i^{te}, relative to x_i^{tr} might contribute to the domain shift [19].

Existence of Uncategorized Instances: The closed space assumption of traditional ML algorithms certainly doesn't hold true in the open world where unseen situations can emerge unexpectedly [20]. It is an inherent fact that the test set may contain some classes that are not present in the training set [21]. Apparently, the model may experience a shift in the semantics of the representation it has learned from the training set as a result of the appearance of these unseen instances. For example, if a binary classifier trained with categories of dog and cat suddenly sees a fox, whilst the covariates of dog and fox might have some correlation, they represent entirely different semantics.

3 Formalizing Distribution Shifts

In this section, we will formalize different distribution shifts by adhering to the official definition of the topic presented in [2] and building upon it in terms of addressing the problem. Abbreviations used in the paper are enlisted in Table. 1.

3.1 Preliminaries and Definitions

Let $\{(x_1^{tr}, y_1^{tr}), (x_2^{tr}, y_2^{tr}), ..., (x_n^{tr}, y_n^{tr})\}$ be the labelled training data sampled from distribution $\mathcal{D}^{tr}(X, Y)$, where $x_i^{tr} \in X$, and $y_i^{tr} \in Y$ represent the i^{th} sample and the associated label, respectively. Similarly, let $\{(x_1^{te}, y_1^{te}), (x_2^{te}, y_2^{te}), ..., (x_m^{te}, y_m^{te})\}$ be the test data sampled from a test distribution $\mathcal{D}^{te}(X, Y)$. Let, $P^{tr}(X)$, $P^{tr}(Y)$ be the marginal distributions and let $P^{tr}(Y|X)$, and $P^{te}(Y|X)$ be the conditional distributions for the training and test data respectively. Based on this, we define the following:

Definition 1: (No-Shift) *The test data is said to be not shifted (in other words in* ***in-distribution (ID)*** *with the training data) when $P^{tr}(X, Y) = P^{te}(X, Y)$. In this scenario, the statistical properties of the data (both the marginal and conditional distribution of the input variables) are assumed to be same between the training and testing phases. i.e., $P^{tr}(X) = P^{te}(X)$ and $P^{tr}(Y|X) = P^{te}(Y|X)$.*

If $L(f(x), y)$ is the loss function for some particular pair of inputs X, and outputs Y, we define following:

Definition 1.1 (True Risk): *is mathematically defined as*:

$$R_{true}(f) = \mathop{\mathbb{E}}_{p_{true}} [L(f(X), Y)] \tag{1}$$

where p_{true} is the true distribution over the x, and y which is unknown.

Definition 1.2 (Empirical Risk): *is mathematically defined as*:

$$R_{emp}(f) = \mathop{\mathbb{E}}_{p_{emp}} [L(f(X), Y)] = \frac{1}{n} \sum_{i=1}^{n} L(f(X_i), Y_i) \tag{2}$$

where p_{emp} is the sampled distribution which consists of limited number of samples quantitatively smaller than p_{true} and can lie in different regions of sample space of p_{true}. Under no-shift condition, which is a fundamental ground of i.i.d hypothesis, it is generally assumed that the empirical risk minimization (ERM) [22] leads to consistent generalization. Since $\mathcal{D}_{tr}$ is the representative of $\mathcal{D}_{te}$, the model trained using ERM will perform similarly on the test data. With large enough training samples, p_{emp} approximates p_{true} well and empirical risk converges to the true risk i.e., $R_{emp}(f) \rightarrow R_{true}(f)$. This renders that for any $f \in \mathcal{F}$, minimizing the empirical risk minimization performs similarly to true risk minimization leading to optimal prediction performance.

$$\arg\min_f R_{emp}(f) \approx \arg\min_f R_{true}(f) \tag{3}$$

Conversely, the test data is said to be Out-Of-Distribution (OOD) when $\mathcal{D}^{tr}(X, Y) \neq \mathcal{D}^{te}(X, Y)$. In this scenario, the model trained with ERM will perform poorly as $R_{emp}(f)$ does not accurately reflect $R_{true}(f)$, and $p_{true} \neq p_{emp}$. Under the OOD framework, we define two independent distribution shifts as follows:

Definition 2: (OOD with Covariate Shift) *The test data is in OOD with covariate shift, when it is subjected to change of distribution in feature space i.e., $P^{tr}(X) \neq P^{te}(X)$, but, the conditional distribution of the target given the input remains unchanged i.e., $P^{tr}(Y|X) = P^{te}(Y|X)$.*

As can be seen from Fig. 1, the training data contains images of dog sitting front of a grass, whereas in the test space (highlighted in blue) the dog appears to be wearing a raincoat in front of a dark background. Here, although the conditional distributions (labels of training and test) should remain same, the features representing the respective labels are distinct. In such cases, based on the inductive biases, the learning algorithm may abruptly fail to correctly classify the test samples.

Definition 3: (OOD with Semantic Shift) *The test data is said to be in semantic shift with the training data when their relationship between the input and the target variables*

change i.e., $P^{tr}(Y|X) \neq P^{te}(Y|X)$. This type of shift can occur regardless of whether the margin distributions of the input variables change i.e., $P^{tr}(X) \neq P^{te}(X)$ or remains the same $P^{tr}(X) = P^{te}(X)$.

In Fig. 1, although the test samples highlighted in red share visual similarity to the samples of $P^{tr}(X)$ (fox might look like a dog, and a cub might look like a cat), the conditional distribution of these samples entirely differ. In this example, the features $P^{tr}(X)$ and $= P^{te}(X)$ share some similarity, while the semantic shift might occur due to entirely different features as well. For instance, a test sample being any non-lookalike object to any of the training set.

3.2 Real-world Shift Dynamics

In practice, during the production phase, machine learning models are exposed to test data that does not strictly adhere to the training distribution. In fact, the test data can be shifted into one of the two prevalent shifts: covariate and semantic whose distributions are denoted as $\mathcal{D}_C^{te}$, and $\mathcal{D}_S^{te}$ respectively. Therefore, mathematically we can say:

$$\{\mathcal{D}_C^{te}, \mathcal{D}_S^{te}\} \subset \mathcal{D}^{te}(X, Y) \tag{4}$$

Now that the i.i.d assumption is routinely violated, ERM does not account for the changes in distribution, leading to suboptimal generalization and prediction performance. Intuitively, when a data is drawn from the sample space of test dataset, then it can experience either experience no shift or be shifted in covariate space or semantic space. In the face of distribution shifts, it is highly important for machine learning models to move beyond the ERM framework to handle the OOD inputs. For instance, for a test sample x_m^{te} drawn from $\mathcal{D}^{te}$, the objective should be minimizing the OOD risk R_{OOD}, which equates to minimizing risk for covariate, R_{cov} and semantic shift, R_{sem} simultaneously.

$$\arg\min_f R_{OOD}(f) = \arg\min_f R_{cov}(f) + \arg\min_f R_{sem}(f) \tag{5}$$

4 Distribution Shifts Mitigation Strategies

In this section, we discuss different approaches employed in the literature to handle different types of distribution shifts. Specifically, we review the modelling strategies in two perspectives; the first one being related to the feature space shift (covariate shift), and the other one associated with the shift in concepts (semantic shift). For each category, we review several approaches by focusing on their corresponding representative works in terms of types of problem domains such as computer vision and natural language processing. Moreover, a global picture of how these approaches are linked to solving a common problem of data distribution shift is discussed.

4.1 Review Structure of the Paper

Our study offers a comprehensive survey of existing methodologies for mitigating data distribution shifts. The paper is structured to guide readers systematically, ensuring optimal comprehension. We discuss, compare, and report benchmark results for each mitigation strategy presented in the taxonomy illustrated in Fig. 2. Initially, we provide a systematic comparison of methodologies, as depicted in Tables 2a and 2b. This comparison is conducted from a fundamental perspective, evaluating different criteria pertinent to each shift type. Subsequently, we present schematic diagrams of each mitigation methodology in Figs. 3 and 4, illustrating the operational mechanisms of these methods concerning data-point/sample classification and decision boundary establishment. Thirdly, we report benchmark results of methodologies across covariate and semantic shifts in Figs. 5 and 6. Here, we present a comparative overview of the highest reported performance metrics (e.g., accuracy, AUROC) for each method on the target dataset, as documented in their respective papers. All values are drawn directly from the original sources, without re-evaluation. Furthermore, to enhance practical guidance for the readers, we also discuss most of these methods in a separate tables (Tables. 3, and 5) where we point out their core working strategy, best use case, and potential limitations. Lastly, through Tables 4, and 6, we provide a comprehensive report of latest research in several applied domains, highlighting corresponding shift handling mechanisms, along with their core technological synopsis.

4.2 Covariate/Feature Shifts

The phenomenon where the distribution of input features (or covariates) in the training data diverges from that in the test data test while the conditional distribution of the targets given the inputs remains unchanged is known as distribution shift in the feature space [23]. This shift can be particularly troublesome in real-world situations where the context or environment in which models are used changes over time or is not the same as the one in which they were trained. Neglecting these changes may result in less than ideal model performance or even model failure.

As investigated by [24], the models trained on ERM often use a cheating way to perform classification, by learning spurious features from the training data which holds no stable properties of the sample. Under covariate shift, often models are very likely to pick up these spurious correlation while missing out the robust features that has causal relationship with the output labels [25]. This inherent fact often gives rise to poor generalization on the new data that are sampled without such biases.

4.2.1 Transfer Learning

Transfer learnings are often used from the data sufficient source task to complement the similar but non-identical target task with limited training samples [88]. The objective of transfer learning technique is to reduce the amount of new labeled data required in the target domain, and possibly avoid the cost of collecting an entire new labeled training data [89]. In the context of transfer learning, a domain is defined by its feature space and its marginal probability distribution while a task is characterized by its label space and an associated objective predictive function. *Transductive transfer learning* [90] explains the phenomenon similar to that of Domain Adaptation, where there is shift in the domains

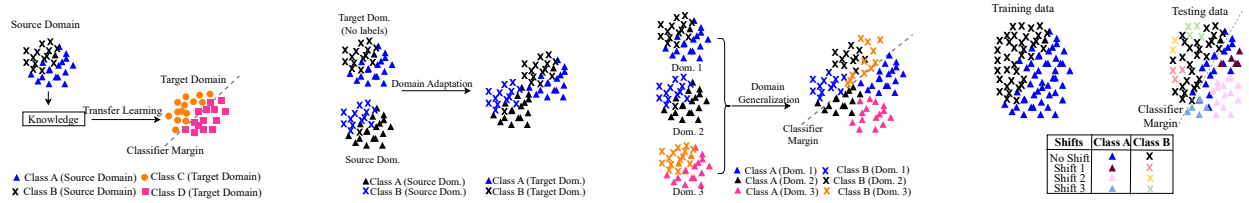
Table 2: Systematic Comparison of Different Types of Shift Mitigation Strategies

(a) Covariate Shift

Criterion	Transfer Learning	Domain Adaptation	Domain Generalization
Distribution Assumption	$P^{te}(X)$ is different but related to $P^{tr}(X)$	$P^{te}(X)$ is shifted but has overlap with $P^{tr}(X)$	$P^{te}(X)$ is entirely unknown
Data Access	Access to labeled $P^{te}(X)$	Access to unlabeled $P^{te}(X)$	No access to $P^{te}(X)$
Learning Method	Pretraining and fine-tuning	Source training and adaptation	Contrastive learning, IRM, Feature disentanglement learning
Testing Strategy	Evaluate on fine-tuned target	Compare performance before and after adaptation	Measure zero-shot generalization
Model Selection	Best pre-trained model for fine-tuning	Best adaptation method per domain pair	Best generalization across domains
Regularization Level	Minimal	High (to prevent overfitting to source domains)	Very high (for domain-invariant learning)
Generalization Capability	Poor	Moderate	Strong

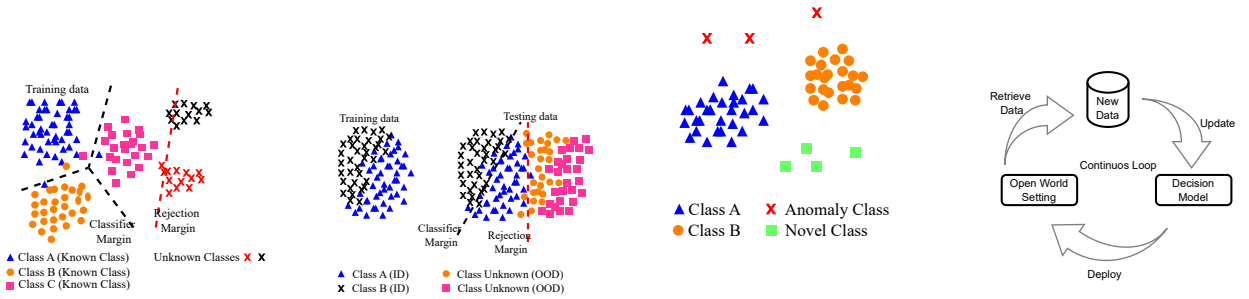
(b) Semantic Shift

Criterion	Anomaly Detection	OOD Detection	OSR	CL
Core Objective	Detecting rare, extreme outliers	Detect entirely different distributions	Balance ID classification and unknown rejections	Learning with evolving data over time
Training Data Assumption	Only ID	OOD data may be available for regularization	ID and OOD data	ID and new OOD classes evolved over time
Test Data Assumption	Rare anomalies that share some features with ID	Samples from entirely different dataset or domain	Samples from ID and OOD distributions	Samples from old ID and newly evolved OOD
Data Availability	Labeled ID, unlabeled anomalies	Labeled ID, unlabeled or uniformly labeled OOD	Labeled ID and unlabeled OOD	Labeled past ID, unlabeled emerging OOD



(a) Transfer Learning (b) Domain Adaptation (c) Domain Generalization (d) OOD Generalization

Fig. 3: Schematic representation of different mitigation approaches for handling Covariate Shift.



(a) Open Set Recognition (b) OOD Detection (c) Anomaly/ Novelty Detection (d) Continual Learning

Fig. 4: Schematic representation of different mitigation approaches for handling Covariate Shift.

of training and test sets without the task being changed. In this kind of setting, it is possible that either the feature spaces are different or the marginal probability distributions of the input data are different [91].

In the past, transfer learning approaches considered specific parts of the model to be carried over between tasks [92], [93], until recently where large cohort of researches [94], [95] focused on the problem of data distribution changes, especially relating to the covariate shift. A study in [96] actually carried out an investigation to answer what knowledge is being transferred from the source domain to the target domain in the process of transfer learning. This study offered novel tools and analysis approach to identify factors that con-

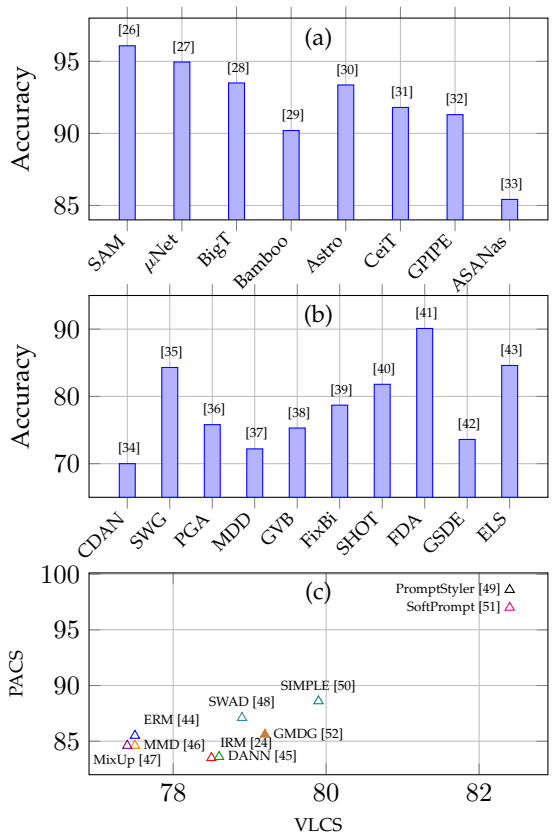


Fig. 5: Results overview of existing popular benchmarks in mitigating Covariate Shift. (a) Transfer Learning accuracies on CIFAR-100 datasets, (b) Domain Adaptation accuracies on OfficeHome dataset, and (c) Domain Generalization accuracies on PACS and VLCS datasets.

tribute to successful transfer and pinpoint the network components responsible for it. [97] delved into the context of the Unsupervised Transfer Learning Challenge, highlighting the benefits of unsupervised pre-training of representations and demonstrating how it can be leveraged in situations where the focus is on generalizing to the instances that originate from a different distribution than the training set. Other variants have extended the fundamental idea of this field into adaptive [98] transfer learning that deals with learning an accurate model using tiny amount of new data, and online transfer learning [99] which makes an assumption of training data in the new domain arriving sequentially.

A TL algorithm to handle both support and model shift was studied in [100]. In this algorithm, the shift handling is performed by transforming both features, and labels of the input by a location-scale shift allowing more flexible transformations. In [101], a conditional shift regression task was studied using deep transfer learning for machine health monitoring in industrial application. The authors specifically proposed a hybrid loss function with achieve two objectives; reducing the prediction error, and preserving global characteristics of conditional distribution dominated by target data. A similar study in TL for regression under conditional shift was conducted in [102] where they considered a special case of source and target domains sharing same margin distributions but non-identical conditional probability distributions. The authors proposed a framework for TL based on fuzzy residual that can learn the target model in a model agnostic way without neglecting the properties of the source data.

The literature in the use case of transfer learning has been substantial, and is not plausible to cover all of them in detail. Therefore, we guide our readers to other applied studies that relays specific use case of transfer learning in reinforcement learning [103], medical image analysis [104], machinery fault diagnosis [105], sentiment analysis [106], intrusion detection systems [107].

4.2.2 *Domain Adaptation*

The statistical attributes of data from any domain might undergo transformations over time, or newly acquired samples could accumulate from diverse sources, leading to what is known as domain shift. When there is a misalignment between the distributions of training and test data, the performance of the trained model is prone to deteriorate upon application to the test data. Domain adaptation (DA) represents a specific subset of transfer learning where labeled data from one or multiple pertinent source domains is leveraged to perform tasks in a distinct target domain [108]. The principal objective of domain adaptation is to learn a model using labeled data from the source domain that can generalize well to the target domain by minimizing the disparities between the domain distributions [109]. There has been numerous study in the long line of literature in the field of supervised DA [110], [111], [112], [113], semi-supervised DA [114], [115], [116] and unsupervised DA [45], [117], [118], [119] attempting to solve the non-trivial task of adapting the source trained model into the target domain in ML systems. Specifically, in this section we try to explain the DA techniques specifically curated for dealing with covariate distribution shift. While our taxonomy groups methods according to their primary design focus, e.g., DA methods under covariate shift it is important to acknowledge that many modern DA variants inherently address aspects of semantic shift as well. Variants such as open-set DA [120], partial DA [121], universal DA [122], and class-incremental DA [123] explicitly account for label space mismatches, including scenarios where the target domain contains unseen or partially overlapping classes. These approaches extend beyond the classical covariate shift assumption (i.e., $P(X)$ changes while $P(Y|X)$ remains fixed) and instead operate in regimes where the conditional distribution $P(Y|X)$ itself shifts, which is central to semantic shift. However, these methods typically address semantic shift arising from label space divergence, rather than deeper semantic reinterpretations of same label.

In [124], researchers have considered a DA study in handling two types of distribution shift; one being the distribution of the covariates, and other the conditional distribution of the target data given cross domain covariate shift. To handle such shifts, the study proposed approaches based on kernel mean embedding of distributions (conditional and marginal), empirically verifying their theoretical claims with experiments on real world problems. Another research studied the DA problem under open set label shift where label distribution can change unexpectedly as well as novel concepts

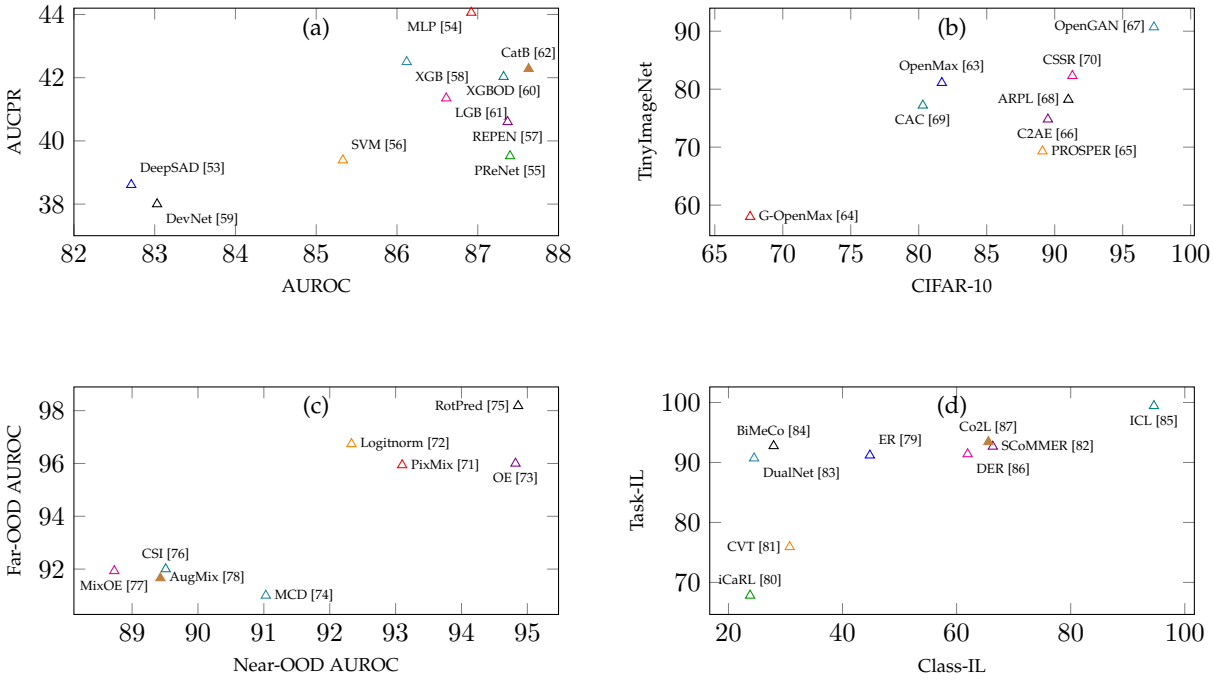


Fig. 6: Results overview of existing popular benchmarks in mitigating Concept Shift. (a) Plot of AUROC vs AUCPR values of existing AD approaches. (b) AUROC plot of OSR methods against two popular benchmark datasets: CIFAR-10 and TinyImageNet. (c) Near and Far OOD AUROC plot of popular OOD detection benchmarks. (d) Accuracy plot of Task-Incremental Learning vs Class-Incremental Learning for popular CL benchmarks.

can appear during deployment [125]. They proposed learning a target classifier, termed as Positive-Unlabeled (PU) learning where the learners objective is to estimate the target label distribution including that of newly introduced classes as well. With rigorous experiments across several vision, medical, and language benchmark datasets, a well-posed problem was offered with significant improvement in the target domain accuracy. Similarly, a large-scale benchmark was introduced in [126] that consisted of over 500 distribution shift pairs across language, vision, tabular datasets. These distribution shifts focused not only on class-conditional shifts but also the label marginal shifts.

4.2.3 *Domain Generalization*

In practical scenarios, it's infeasible to collect training data across every conceivable domain. The capacity of a model to extrapolate from familiar domains to unfamiliar ones is paramount. Domain generalization addresses the intricate task of educating a model using data from one or several source domains so that it can adeptly generalize to novel, unseen target domains that share the same label space. Very often in the literature [127], [44], [128] this technique is used to in conjunction with out of distribution (OOD) generalization [129]. The only difference being the former employs multiple training datasets from different domain for model training purpose [130]. Nevertheless, this is achieved without the luxury of accessing any data from the target domain during the training phase. While the OOD generalization is a more generic term, both share the same objective of generalizing well on unseen domain by capturing domain-agnostic representations.

In the realm of machine learning, the generalization aptitude of a model is frequently contingent upon the volume and heterogeneity of the training dataset. When confronted with a constrained dataset, data augmentation emerges as one of the most cost-effective and straightforward strategies to proliferate samples, thereby bolstering the model's generalizability. The primary aim of data augmentation-driven techniques is to amplify the variance within the existing training dataset by employing diverse data manipulation methodologies. Concurrently, this process also augments the overall volume of the dataset. One of the DG study in [131] aimed to improve the performance estimation of the model in the presence of distributional shift without supervision. They used a set of domain-invariant representations as a proxy model for an unknown true target labels where the accuracy of the resulting risk estimates depended on the target risk of that model. The study addressed the generalization of range-invariant representations and showed that the complexity of the latent representation has a significant impact on target risk. Empirically, their method facilitated self-tuning of the DA models while accurately estimating the target error of a given model under distributional shifts. Another empirical paper [132] studied the problem of graph OOD generalization by evaluating eight different datasets representing different types of distributional shifts on graphs. These datasets were used to perform a comprehensive empirical evaluation of popular DG algorithms, graph expansion methods and GNN models. They came up with an interesting deduction that most DG algorithms did not improve OOD generalization performance when confronted with different types of domain shifts on the graph. Instead they discovered that the optimal combination of advanced GNN models and robust graph expansion methods can effectively achieve

state-of-the-art performance in graph OOD generalisation problem.

Unlike the straightforward application of principle of invariance in images, identifying invariant features within graph data is inherently challenging. A study in [133] addressed this challenge of applying invariance principles to graph data under distribution shifts. In particular, they proposed a causality inspired invariant graph learning to ensure OOD generalization on graphs. The authors claim that true OOD generalization can be achievable if the focus is shifted towards subgraphs that hold substantial information regarding the causality behind label assignments. The proposed technique involved an information-theoretic objective designed to identify and safeguard these invariant intra-class information, thereby ensuring that the learned subgraph representations are resilient to distribution shifts.

4.3 Concept/ Semantic Shifts

[154] defines concept as a function learned by an algorithm that maps input values to their corresponding output values, as defined by a set of training examples. Concept shift, or drift in some literature [155] occurs when the posterior probabilities of the input and labels change. Due to the complex and dynamic nature of data distribution shift that occur over time, the model can be presented with new concepts (e.g. new categories of objects) at any time [156]. The introduction of new concepts can result in the catastrophic failure of a model due to its reliance on the iid hypothesis for prediction. Therefore, handling these shifts is essential for maintaining the robustness of ML models. Often in computer vision, the change of concept is used interchangeably with change in semantics as it represents the change in features intrinsic to the object [157], [7].

In this section, we outline various methodologies and techniques that have been studied as part of addressing this problem, with either explicit or implicit association with the concept shift.

4.3.1 Open Set Recognition

Traditional classification methods require the system to classify all the test instances into one of the trained classes, disregarding the prevalence of concept shift that takes places in an open world. Instead a robust ML system must constrain its classification criteria within the known paradigm of learned classes whilst rejecting unseen classes that are irrelevant and meaningless to what it has learned. Open set recognition holds two supposition to enhance the robustness of ML systems; one is to accurately classify samples into known categories, and the other is to detect and reject unknown samples [8]. Sometimes also referred as open world machine learning [158], this approach aims to eliminate the risk of mistakenly categorizing an unknown instance into one of the known categories. One of the challenges in formulating OSR method is to optimize the model for accurately estimating the probability of all known classes while maintaining precise recognition of the unknown classes [159]. In this regard, several approaches have been studied in ML research to address this concept shift problem.

In general, majority of DNN's penultimate layer are connected to the Softmax layer which is responsible to produce a probability distribution over the total number of known concepts it is trained on [160], [161]. It was a common technique for handling samples arising from unknown concept by assigning a threshold with an assumption that unknown samples would incur low probability. However, this type of uncertainty thresholding technique was later found to be simply not enough to determine what is unknown because of two reasons. First, the unknown samples are usually known to hold a really large sample space in an open world and generalizing to this large subspace was challenging. Second, the counterfactual, and adversarial images can really fool the model by producing high confidence scores regardless of them being unknown. To cope up with this issues, one of the preliminary study in OSR proposed OpenMax [63] where the existing deep neural network (DNN) was modified by introducing a new layer with the objective of assessing the likelihood of the input belong to an unknown concept. By applying distance normalization process based on extreme-value meta recognition on the activation patterns of DNN's penultimate layer, the rejection probability was determined for unknown images. By doing so the system was able to effectively reject misleading, unknown, and even many adversarial images significantly reducing the obvious error of traditional DNN's in open space. Several consecutive studies were studied aiming to enhance the OSR benchmark results. An interesting OSR research based on sparse representation was studied in [162] which used class reconstruction errors for classification task. The proposed framework is grounded on the principle of Extreme Value Theory (EVT) and unfolds in two main stages. The approach initially models the tail distributions of both matched and non-matched reconstruction errors by employing EVT, thereby transforming the intricate OSR issue into two separate hypothesis testing situations. Afterwards, in the second phase, the method entails calculating the reconstruction errors for a test sample from each category and the confidence scores that originate from the two tail distributions are combined to identify the test sample's actual identity.

4.3.2 Out-of-Distribution Detection

The emergence of out-of-distribution (OOD) detection [163] in the field of deep learning is a response to the common issue of models being overconfident in classifying samples from different semantic distributions in image classification and text categorization tasks [7]. This methodology relies on a scoring function that converts the input into an OOD score, signifying the extent to which the sample is considered differently distributed from that of training data [164]. Although, the separation of ID and OOD data remains a non trivial task, it is arguable that continued research progress in OOD detection requires insights into the fundamental cause and mitigation of model overconfidence on OOD [165].

In [166], the researchers introduced GradNorm, a simple yet effective method that utilizes information from the gradient space to identify OOD inputs. GradNorm specifically utilized vector norm of gradients, which are backpropagated from the Kullback-Leibler divergence between the softmax output and a uniform

Table 3: List of Popular Covariate Shift Mitigation Papers, including their Core Strategy, Best use case, and Limitations.

Type	Reference (Year)	Core Strategy	Best use	Limitation
	SAM [26] (2020)	Minimizes sharpness and loss jointly	Reduces overfitting to source domains	High compute cost; ignores domain structure
	μNet. [27] (2022)	Evolutionary search over modular subnets	Modular transfer across domains	Costly search; limited to predefined modules
TL	BigT [28] (2020)	Transfer of large pre-trained models	Strong baseline under covariate shift	Requires massive compute and data resources
	Bamboo [29] (2022)	Online model selection with bootstrapped risk	Dynamic adaptation under shift	Needs access to unlabeled target data streams
	Astro [30] (2023)	Transformer-based meta-learned initialization	Few-shot adaptation under distribution shift	High model complexity; meta-training required
	CeiT [31] (2021)	Combines CNN token embedding with Vision Transformer	Visual tasks needing inductive bias	Less effective on small data or non-visual domains
	CDAN [34] (2018)	Aligns joint feature–label distributions via adversarial training	Supervised DA with label shift	Sensitive to classifier confidence and adversarial training stability
	SWG [35] (2023)	Aligns Wasserstein geometry between domains	Unsupervised DA with strong structure shift	Computationally intensive; limited scalability
	PGA [36] (2024)	Promotes geometry alignment using pseudo labels and self-training	Unsupervised DA with label imbalance	Relies on pseudo label quality; sensitive to noise
DA	MMD [37] (2020)	Minimizes domain discrepancy via MMD loss	Simple and effective DA baseline	Simple and effective DA baseline
	GVB [38] (2020)	Gradually aligns features via auxiliary boundary loss	DA with class-boundary refinement	Requires careful scheduling and tuning
	FixBi [39] (2021)	Bidirectional self-training with reliable pseudo labels	Class-balanced unsupervised DA	Sensitive to early pseudo label errors
	SHOT [40] (2020)	Source-free DA using feature clustering and pseudo labeling	DA when source data is unavailable	Relies on target structure; tuning is tricky
	ERM [44] (2020)	Minimizes average empirical risk across source domains	Strong baseline for DG	Ignores domain-specific signals and variance
	IRM [24] (2019)	Learns invariant predictors across domains	DG with strong causal assumptions	Hard to optimize; often underperforms in practice
	DANN [45] (2015)	Adversarial feature alignment via domain classifier	Early DA and DG benchmark	May confuse domain-invariant and task-relevant features
DG	MixUp [47] (2017)	Interpolates inputs and labels for regularization	Improves generalization and robustness	May underperform on complex or structured shifts
	SWAD [48] (2021)	Averages weights from flat minima for stable generalization	DG under training instability	Assumes flatness correlates with generalization

Table 4: List of Applied Research Papers Dealing with Covariate Shift Problem

Type	Reference (Year)	Applied Area	Core Technology	Dataset Type	Used Metrics
	Sohn et al. [134] (2023)	Image Synthesis	Generative vision transformers, Prompt tuning	Vision data	FID, LPIPS
	Qian et al. [135] (2023)	Machine Fault Diagnosis	Conditional alignment, I-Softmax loss	Vibration data	Accuracy
TL	Zhu et al. [136] (2023)	Machine Fault Diagnosis	Bayesian semi-supervised TL, MC dropout	Vibration data	RMSE, MAE
	Bierbrauer et al. [137] (2023)	Intrusion Detection	1D CNN, Random Forest	Network traffic	TPR, FPR, TNR
	Zhou et al. [138] (2024)	Intelligent Transport	Federated TL, Siamese NN, Spatio-temporal clustering	Positioning data	RMSE, MAE, MAPE
	Xiao et al. [139] (2024)	Heterogeneous Labels	Random Walk, LSTM, Meta-learning	Image data	Transfer Accuracy
	Li et al. [140] (2024)	Image Classification	Variational NN, Conditional alignment	Synthetic, Multi-domain	Adaptation Accuracy
	Hoyer et al. [141] (2023)	Visual Recognition	Unsupervised DA, Masked image modeling	Segmented images	Accuracy, IoU
	Truong et al. [142] (2023)	Scene Understanding	Conditional structure net, Self-attention	Vision data	Accuracy
DA	Kim et al. [143] (2023)	Text-to-Image	CLIP, Diffusion, Pose filtering	Text-image, 3D data	KID
	Wang et al. [144] (2023)	Text Understanding	Conditional alignment, Optimal transport	Multi-domain image data	Accuracy
	Chen et al. [145] (2023)	Semantic Segmentation	Dual-path translation, ClassMix	Landscape images	Accuracy, IoU
	Hao et al. [146] (2023)	Video-Text Retrieval	Dual alignment, Cross-modal embedding	Video-text pairs	Median ranking
	Ge et al. [147] (2023)	Image Classification	CLIP, Prompt learning, Domain embedding	Text-image pair	Accuracy
	Wang et al. [148] (2023)	Image Classification	Sharpness-aware gradient matching	Image data	Accuracy (ID/OOD)
	Segu et al. [149] (2023)	Image Classification	Batch norm variants, Latent space learning	Image data	Generalization Accuracy
	Chen et al. [150] (2023)	Image Recognition	Federated learning, Adaptive normalization	Image data	Generalization Accuracy
DG	Yu et al. [151] (2023)	Node Classification	Label-invariant augmentation, GNN	Synthetic graph data	Generalization Accuracy
	Zhang et al. [152] (2023)	Image Classification	Bilevel optimization, Adapter layers	Image data	Generalization Accuracy
	Wang et al. [153] (2023)	Data Mining	Conditional independence test, Causal selection	Synthetic, Medical image	RMSE, Avg. error

probability distribution. The key assumption underlying this approach is that the magnitude of gradients is generally greater for ID data than for OOD data. This characteristic makes the gradient magnitude a useful metric for detecting OOD inputs. In another study in [167] the researchers put forward continuously adaptive out-of-distribution (CAOOD) detection framework that was developed with the intention of creating a model that could rapidly adapt to new distributions, especially when there are insufficient ID samples available during deployment. Specifically, the authors devised a meta out-of-distribution learning (MOL) strategy which involved creating a 'learning-to-adapt' diagram that facilitates the initial learning of an effectively initialized OOD detection model during the training phase. During the testing phase, MOL aimed to maintain the OOD detection's efficiency across varying distributions by allowing for swift adaptation to new distributions through minimal adjustments.

One of the core difficulties in OOD detection is that OOD inputs can be extremely diverse, and without any assumptions, detecting anything that's not ID is provably impossible. In other words, if we place no restrictions on what the OOD data could be, no finite training procedure can guarantee detection of every possible OOD input. Intuitively, an algorithm that works well for one type of unseen data can always be fooled by another type, unless we have some prior knowledge or constraints. Therefore, theoretical analyses of OOD detection [168], [169], [170] introduce explicit assumptions or models of the data to make the problem tractable. A common assumption is that the ID and OOD distributions are sufficiently distinct in some feature space (for example, they may have disjoint support or minimal overlap). If OOD examples can occupy the same feature regions as ID ones, then no detector can perfectly separate them, and therefore any decision rule will make errors when ID and OOD data overlap.

Many recent OOD works [171], [172], [173] follow the idea of adopting auxiliary dataset to regularize the model for improving distinctness between ID and OOD data. These techniques are based on the assumption that the auxiliary datasets represent real OOD data, and, utilizing them as a known priori while training along with ID data can actually aid in generalizing to detect unseen distributions. The benchmark OOD study in using auxiliary OOD data is Outlier Exposure [73]. In this paper, authors used a set of outliers that are disjoint from the real OOD test set are used to train the model s to discover signals and learning effective heuristics to detected whether the input belongs to either ID or OOD. While most of these studies follow random sampling of the outliers, other works have considered mining outliers through adversarial training [174], posterior sampling [175], or using leveragin wild mixture data of ID and OOD [176].

4.3.3 Anomaly/Novelty Detection

It becomes crucial that a machine learning system be able to distinguish between known and unknown object information during testing, since it is not plausible to train on all potential objects the system is likely to encounter in the real world. In other words, it is crucial that the robust ML systems must have the ability to identify a set of unlabeled instances that significantly differ from the training dataset. Anomaly and novelty detection, often used in tandem throughout the literature, deal with this problem of recognizing anomalous and novel concepts in the system [177]. Very subtle difference persists among them, as in the former tries to exclusively find negative samples or pecularities, while the latter focuses on discovering novel concepts that needs to be incorporated into the decision model. Nevertheless, both are concerned over finding the OODness in data where the training samples experiences an abrupt change in concept. It is also to be noted that for any classification system, particularly in data streams, two phenomenon can co-exists: concept evolution, which refers to the emergence of new classes, and concept drift, where the known concept can change over time [178].

In paper [179], the authors acknowledged that novelty class in general is either often missing in training, sampled inadequately, or poorly defined, thereby making one-class classifiers a suitable solution for such difficulties. Despite this issue, they proposed an end-to-end architecture specifically for one-class classification, inspired by the success of GANs in training deep models under unsupervised and semi-supervised frameworks. The architecture consisted of two deep networks that are trained together but in opposition, with one network functioning as the novelty detector while the other reinforcing the inlier samples and distorting the outliers. The core idea behind this approach was that the separability between the enhanced inliers and the distorted outliers is substantially greater than when making decisions based on the original samples. Another study [180] aimed to tackle the ND challenges, one of which involved recognizing deviations from a typical model of regularity. This task is made difficult by the unpredictable and often undetectable nature of new concepts during training. To address this challenge, the authors created a comprehensive framework that combined a deep AE with a parametric density estimator to learn the underlying probability distribution of latent representations in an autoregressive manner. By optimizing a maximum likelihood objective along with normal sample reconstruction, their approach effectively regularized the task by minimizing the differential entropy of the latent vectors.

A prominent technique for self-supervised representation learning is to semantically contrast similar and dissimilar sample pairs [181]. Considering this, studies such as [182], [183], [184], [76] have utilized contrastive learning (CL) framework for realizing AD task. One of the impressive works by [182] exploited task agnostic way of using CL in an AD problem where the agreement between differently augmented views of the same image is maximized while repelling with the others in the same batch. By doing so, this method was able to obtain effective representation of each data sample while robustly clustering each class without the necessity of human supervision or labelling. A following work in [183], introduced a task-specific variant of CL, termed masked contrastive learning (MCL). Specifically, they

unveiled an inference method called self-ensemble inference, designed to enhance performance by exploiting the skills acquired through auxiliary self-supervision tasks. The primary insight of their research was that forming dense clusters, without the necessity for fine-tuning yet preserving individual representations leads to the development of more meaningful visual representations. This approach deviated from the traditional 'pre-train then tune' paradigm practiced by [182] and led to effective identification of anomalous data. This study outstood among CL based methods because of its idea of generating a mask that properly adjusted the repelling ratio while taking into account the class labels present in the batch.

4.3.4 Continual Learning

Despite the fact that human learning has developed to excel in environments that are constantly changing and evolving, current machine learning systems are only able to perform effectively when presented with well-balanced and homogeneous data. When faced with data that is otherwise, these models often struggle and not only experience a significant decline in performance but also exhibit a catastrophic forgetting phenomenon on previously learned tasks [185]. A study in [186] has termed these phenomenon as interferences that are explicitly caused by the changes in the data distribution or in the learning criterion. Continual learning (CoL), also referred to as lifelong learning [187] or incremental learning [188] share the mutual goal of developing ML algorithms that do not stop learning, but instead keep model parameters updated to accumulate knowledge over time [189]. Modern dynamic data sources can be affected by shifts that can happen over time, (*concept drift*) where the property of some or all classes might abruptly change. Therefore this calls for the demand of CoL models that can effectively adapt to concept drift scenarios in any data stream mining tasks.

A study in [190] has set down requirements for CoL such that; a learning method that continually improves should not experience catastrophic forgetting, meaning it should maintain its ability to perform well on previously learned tasks. Additionally, it should be capable of learning new tasks while leveraging knowledge gained from earlier tasks, demonstrating positive forward transfer for faster learning and improved final performance. The method should also be scalable, able to be trained on a large number of tasks. Furthermore, it should allow for positive backward transfer, meaning that learning a new task can lead to immediate improved performance on previous tasks that are similar or relevant. Lastly, the method should be able to learn without requiring task labels and ideally be applicable in the absence of clear task boundaries. CoL has traditionally navigated data-constrained scenarios within a supervised framework, where batches of labeled samples were sequentially introduced to the network, enabling it to incrementally assimilate new information while retaining previously acquired knowledge. The research in [191] proposed a method for unsupervised CoL associating unsupervised domain adaptation (UDA) and CoL paradigms. This study addressed the challenge posed by a gradually evolving target domain, segmented into multiple sequential batches, necessitating the model to continuously adapt to the progressively changing data stream without supervision. To address this challenge, they introduced a source-free approach utilizing episodic memory replay coupled with buffer management. Furthermore, a contrastive loss component was integrated to enhance the alignment between the buffer samples and the ongoing flow of batches, aiming to refine the model's adaptability and retention capabilities in the face of evolving datasets.

A CoL system is expected to maintain both plasticity, the acquisition of new knowledge, and stability, the preservation of old knowledge. Catastrophic forgetting represents a failure in stability, where new experiences overshadow previous ones. The authors of [79] utilized replay of past experiences motivated from the approach in neuroscience to mitigate such forgetting. In this work, the authors introduced, a replay-based method designed to significantly diminish catastrophic forgetting within multi-task reinforcement learning environments. The proposed method incorporated off-policy learning and behavioral cloning from replay to strengthen stability, while also employing on-policy learning to ensure the maintenance of plasticity. The paper demonstrated significant performance in mitigating forgetting while referring that the method to be extremely less sophisticated with no requirements of knowledge of tasks being learned. Interestingly, a study was proposed in [192] with counterarguments that experience replay leads to significant overlap between the representations of newly added and previous classes, resulting in highly disruptive parameter updates. This study proposed insights to reduce the abrupt change in data representations that occurs when unobserved classes emerge in the data stream. Based on empirical analysis, a new method was proposed to address this problem by protecting the learned representations from drastic adaptations required to accommodate new classes. They showed that using an asymmetric update rule, which encourages new classes to adapt to older ones, is more effective, particularly at task boundaries where significant forgetting typically occurs.

5 Closely Related Topics

5.1 Runtime Monitoring

Critical software systems based on ML, such as autonomous vehicles, may exhibit abnormal behavior suddenly, severely, and unpredictably while in operation [220]. Ensuring the safety of these systems is extremely challenging during the design phase. Runtime monitoring is a technique that focuses on monitoring the safety of operation by following the current input and raising an alarm when the safety might be violated, rather than checking the correctness of all inputs universally [221]. Such techniques aim to identify unsafe predictions for a given ML model and discard them before they can lead to any catastrophic repercussions. This is usually accomplished by identifying the inputs that are different from the training data [222], [223].

5.2 Open World Recognition

Open World Recognition (OWR) [224], [225] posits that newly discovered categories ought to be continuously identified and subsequently incorporated into the

Table 5: List of Popular Semantic Shift Mitigation Papers, including their Core Strategy, Best use case, and Limitations.

Type	Reference (Year)	Core Strategy	Best use	Limitation
	DeepSAD [53] (2019)	Learns compact representation for normal data via semi-supervised loss	ND with few labeled anomalies	Struggles with complex or overlapping classes
	PReNet [55] (2023)	Progressive refinement of prediction via contrastive pretext tasks	Fine-grained AD in semantic space	Requires careful task design and tuning
	SVM [56] (2003)	Maximizes margin between classes for robust separation	Classical baseline for AD	Struggles with high-dimensional or nonlinear data
	REPEN [57] (2018)	Learns distance-aware embeddings for anomaly ranking	Unsupervised outlier detection in high dimensions	Requires sampling strategy; less effective on structured semantics
	XGBOD [60] (2018)	Boosted ensemble of unsupervised detectors and selected features	Hybrid AD with tabular data	Requires good base detectors; less generalizable
	OpenMax [63] (2016)	Calibrates softmax scores using extreme value theory	OSR with known–unknown separation	Assumes well-structured class distributions
	G-OpenMax [64] (2017)	Enhances OpenMax by generating unknowns with GAN	Detecting unknowns in controlled visual domains	Requires high-quality generative models
	PROSPER [65] (2021)	Introduces prototype-based unknown classifiers with soft likelihood	Large-scale OSR	Sensitive to prototype quality and feature overlap
OSR	C2AE [66] (2021)	Combines class-conditioned autoencoders with discriminative embedding	Open-set detection with reconstruction cues	Struggles with visually similar unknown classes
	ARPL [68] (2021)	Learns reciprocal points and margin-based embedding with adversarial training	Open-set robustness with semantic structure	Sensitive to margin settings; adversarial overhead
	CSSR [70] (2022)	Reconstructs semantic features via class-specific decoders	Open-set recognition with class-aware reconstruction	Requires reliable class prototypes; complex training
	PixMix [71] (2022)	Augments data by mixing with unrelated images	OOD detection under severe distribution shifts	May hurt ID performance; lacks semantic control
	LogitNorm [72] (2022)	Normalizes logits to calibrate softmax scores	OOD detection with pretrained classifiers	May degrade accuracy if misconfigured
	OE [73] (2018)	Trains model with auxiliary outlier data	Improves OOD detection during training	Depends on quality and diversity of outliers
OOD-D	MCD [74] (2024)	Uses ensemble of classifiers with divergence-based OOD scoring	OOD detection with uncertainty estimation	Computationally heavier; sensitive to ensemble diversity
	RotPred [75] (2019)	Uses self-supervised rotation prediction as an auxiliary task	OOD detection with limited labeled data	Less effective on non-visual or abstract inputs
	MixOE [77] (2023)	Mixes ID and OOD samples for contrastive supervision	Fine-grained OOD detection during training	Requires labeled OOD or curated outliers
	AugMix [78] (2019)	Applies diverse and stochastic augmentations with consistency loss	Robust OOD detection and improved generalization	May not capture semantic anomalies
	ER [79] (2019)	Stores and replays past samples during training	Simple and effective continual learning	Memory overhead; prone to sampling bias
	iCaRL [80] (2017)	Combines rehearsal with nearest-mean-of-exemplars classification	Class-incremental learning	Requires exemplar storage; suffers from imbalance
	CVT [81] (2022)	Online distillation with continual vision transformer adaptation	Continual learning in vision tasks	Requires careful update strategy; compute-heavy
CoL	SCoMMER [82] (2023)	Sparse memory retrieval with modular experts	Lifelong learning with minimal forgetting	Complex memory management; tuning expert modules
	DualNet [83] (2021)	Maintains plastic and stable branches for dual-memory learning	Stability–plasticity trade-off in continual learning	Increased model size and training complexity
	BiMeCO [84] (2023)	Bilateral memory consolidation with contrastive objectives	Continual learning with domain shifts	Sensitive to memory balancing and contrastive tuning

recognition process for practical applications. In fact, the system must be capable of recognizing objects and assigning them to existing classes, as well as labeling items as unknowns depending on how these objects are distribution-shifted from the learned data. If there are novel instances unknown to the trained model, then they must be gathered and labeled, for example, by humans. Once there is a sufficient quantity of labeled unknowns for class learning, the system must incrementally learn and expand the multi-class classifier, thus rendering each new class ”known” to the system. Open World recognition extends beyond mere robustness to unidentified classes and instead aims to create a scalable system that can adapt and learn in an open world amidst the challenging distribution shift phenomenon.

5.3 Zero-shot Learning

Zero-shot Learning (ZSL) [226] refers to a method of training a model to classify objects from unseen classes by leveraging knowledge from seen classes through the use of semantic information. Usually, this information is provided in the form of high-dimensional vectors that encompass the names of both the seen and unseen classes. The technique of ZSL essentially bridges the gap between the two types of classes by utilizing semantic information. This approach to learning can be compared to how a human recognizes a new object by assessing the likelihood of its descriptions aligning with previously acquired knowledge. A primal example of this is recognizing a zebra as a horse with black and white stripes, in regards that one has previously encountered horses. We can find several studies of ZSL in the literature for generalized zero shot learning [227], and [228], zero shot domain generalization [229]. We guide our readers attention to comprehensive survey papers in ZSL in [230], and [231].

Table 6: List of Applied Research Papers Dealing with Covariate Shift Problem

Type	Reference (Year)	Applied Area	Core Technology	Dataset Type	Used Metrics
	Bashari et al. [193] (2024)	Intrusion Detection	Conformal inference, RF, SVMs	Synthetic, Network traffic	Power, FDR, Variance
	Zhu et al. [194] (2023)	Spam Detection	Margin theory, Multi-class novelty detection, SVM	Remote Sensing	AUC, Error
	Liu et al. [195] (2023)	Industrial Defect	Pyramid deformation, Diversity-based detection	Industrial, Surveillance video	AUC
AD/ND	Xu et al. [196] (2023)	Cyber Intrusion	SMOTE, Multi-class classification	Network traffic	Accuracy, NMI, F1
	Chen et al. [197] (2023)	Video Anomaly	Feature amplification, Magnitude contrastive loss	Video data	AUC, AP
	Xie et al. [198] (2023)	Industrial Anomaly	Few-shot, Graph representation	Industrial images	AUROC
	Wang et al. [199] (2023)	LLM	Adversarial + OOD robustness, Zero-shot	Product reviews	Adversarial/OOD robustness
	Graham et al. [200] (2023)	Image Recon.	Diffusion models, Denoising Autoencoders	Medical images	AUC
	Wu et al. [201] (2023)	Node Classification	GNNs, Energy Function	Image data	AUROC, AUPR, FPR, ID-Acc
OOD-D	Zhang et al. [77] (2023)	Image Classification	Mixup, Outlier Exposure	Image data	Accuracy, TNR95
	Wilson et al. [202] (2023)	Image Classif.	Hyperdimensional computing, Gram detectors	Image data	AUROC, FPR95, Detection error, F1
	Song et al. [203] (2023)	Text Understanding	Text clustering, Neuron activation	Text data	User rating
	Li et al. [204] (2023)	Image Classification	Vision Transformer, One-class finetuning	Image data	AUROC
	Mereau et al. [205] (2024)	Text Classification	Max softmax, Mahalanobis KL, Rank-weighted depth	Movie reviews	AUROC, AUPR
	Liu et al. [206] (2023)	Text Recognition	Label-to-prototype, Zero-shot	Chinese text images	Open-set accuracy
	Liu et al. [207] (2023)	Image Classification	Gaussian prototypes, Bayesian inference	Image data	AUROC, Precision, Recall
	Sun et al. [208] (2023)	Image Recognition	Hierarchical Attention, LSTM	Image data	Accuracy, AUROC, OSCR
	Yang et al. [209]	Image Classification	Model attribution, Sample augmentation	Facial images	AUC, OSCR
OSR	Zhang et al. [210] (2023)	Action Recognition	Reconstruction, Discriminative features	Video data	Accuracy (open/closed set)
	Li et al. [211] (2023)	Text-to-Image	Diffusion, Zero-shot gen.	Image-text pair	FID, AP
	Li et al. [212] (2023)	Object Detection	Language-guided query, Modality fusion	Image-text pair	Accuracy, AP
	Soltani et al. [213] (2023)	Intrusion Detection	Deep clustering, SVMs	Network traffic	Accuracy, Misclass. error
	Smith et al. [214] (2023)	Image Classification	Rehearsal-free, Param. regularization	Image data	Accuracy, Forgetting
	Villa et al. [215] (2023)	Video Classification	Multi-modal classifier, Prompting	Action data	Accuracy, BWF
	Raz. et al. [216]	Language Model	Progressive prompting, Embedding reparam.	Online reviews	SuperGLUE, FWT, BWT
CoL	Yuan et al. [217] (2023)	Driving Action	P2P federated learning, IoV	Driver video	Objective, Generalizability
	Yang et al. [218] (2023)	Image Classification	Bayesian GMMs, Incremental learning	Image data	MCR
	Zhu et al. [219] (2023)	Semantic Segment.	MDP, Memory sampling, Graph struct.	Image data	Accuracy, IoU

6 Discussion

6.1 Distribution Shift in Large Language Models (LLMs)

Large Language Models (LLMs) are susceptible to vulnerabilities resulting from distribution shifts. An LLM trained on a specific corpus may exhibit reduced performance when there are alterations in the input language [232], domain [233], or task distribution. Such shifts can significantly impair accuracy and increase perplexity, thereby compromising real-world reliability. In the event of a covariate shift, an LLM may encounter difficulties with unfamiliar vocabulary, styles, or structures, leading to misinterpretations. Empirical studies have demonstrated that even large pre-trained models exhibit a notable decline in performance when evaluated on OOD data. For example, in the WILDS benchmark [234] of real-world shifts, models trained on one domain consistently exhibited substantially lower OOD accuracy compared to ID accuracy.

Covariate shifts primarily affect the recall of knowledge and alignment to input, where the model may fail to recognize entities or idioms it has not previously encountered (e.g., a medical term abbreviation), or it may incorrectly parse syntax when faced with code or XML in the input. In generation tasks, covariate shifts can lead to incoherence or irrelevant continuations. Notably, LLMs demonstrate some resilience to mild covariate shifts due to their extensive training data; however, under severe shifts, such as transitioning to a different language or a highly specialized jargon, performance can degrade abruptly. This is particularly evident in zero-shot settings where the model has not been conditioned on that style [235]. Additionally, concept drift may occur if language usage patterns change such that the model's learned correlations no longer hold (e.g., a word previously indicating negative sentiment is adopted as slang for something positive). Unlike covariate shifts, concept shifts typically alter the decision boundary or generation mapping and may necessitate relearning the task function, such as through fine-tuning on new examples [236]. Given the inevitability of distribution drift in real-world data, a variety of strategies have been developed to maintain or improve LLM performance under shift conditions [237]. Such strategies encompass retraining or fine-tuning the model with new data, implementing on-the-fly adjustments such as prompting or retrieval, and assessing uncertainty.

6.2 Practical Applications: Impact of Distribution Shift

Distribution shifts can profoundly affect a model's performance in practical applications, leading to substantial decrease in the accuracy of downstream tasks. This decline can have serious implications for the decision-

making processes in critical systems such as medical diagnosis, autonomous driving, fault detection, intrusion detection, and adversarial defenses. For example, a medical imaging model trained on data from one hospital may not perform well on scans from another due to one of main reasons including the differences in data acquisition device, patient demographics, or imaging protocols. This phenomenon, known as domain shift, can often lead to misdiagnoses or even serious health consequences of the patient. Likewise, autonomous vehicle models trained in sunny conditions may struggle in rain or snow, as their perception systems rely on visual patterns that change with environmental conditions [238].

In consumer applications such as recommendation systems and spam detection, distribution shifts can result in user dissatisfaction or exploitation. For instance, changes in user behavior over time, such as trends or seasonal interests, can render previously effective recommendation algorithms obsolete, necessitating constant updates [239]. In adversarial contexts such as spam or fraud detection, attackers may deliberately induce concept drift, a type of distribution shift to evade detection by exploiting model vulnerabilities [240]. Even in seemingly stable applications like language models or image classifiers, OOD inputs or subtle distribution shifts can lead to high-confidence but incorrect outputs. A prime example of this can be a chatbot trained on standard internet data may produce inappropriate or biased responses when encountering unfamiliar slang, dialects, or cultural contexts. This is particularly concerning in open-world settings, where inputs may come from unpredictable or evolving sources. Overall, distribution shifts undermine the generalization capabilities of machine learning models and reveal the fragility of systems that are not robust to changes in data distribution. They also complicate model evaluation, as performance metrics on held-out test sets may not accurately reflect real-world reliability.

6.3 Challenges

The distinction between covariate and semantic shifts, while useful for theoretical delineation, may be overly simplistic when addressing practical ML challenges, where these shifts often occur simultaneously and are intertwined. The current categorization - TL, DA, DG/OOD-G for covariate shifts, and OSR, OOD detection, AD/ND, and CoL for semantic shifts - has undoubtedly revolutionized our understanding and capability to tackle each type of shift. Yet, this segregation does not reflect the complexity of real-world applications, where shifts do not present themselves in isolation. For instance, an autonomous driving system may face varying weather conditions (covariate shift) while also encountering new road signs or alterations (semantic shift). Addressing these shifts independently may not be sufficient or efficient for robust performance. While a holistic approach can enhance the algorithm's practical ability to learn from intricate, multifaceted shifts, improving generalization and robustness across diverse situations. In discussing individual methods, we also put forth recently emerging studies that aim to bridge the gap between these two types of shifts, demonstrating the feasibility and efficacy of comprehensive strategies [241], [242], [243], [157]. These pioneering works indicate that ML algorithms can be designed to be inherently adaptive, and detecting while pursuing to handle both covariate and semantic shifts under one cohesive framework.

7 Future Research Directions

Developing ML models that can effectively handle data distribution shifts necessitates coordinated efforts across numerous research areas. In order to drive the development of mechanisms to handle the distribution shifts we bring forward several potential future research directions.

Strong Foundation and Benchmarks: Since the modelling of a unified framework capable of addressing both covariate and semantic shifts simultaneously is of paramount importance, one prospect is to strengthen the current theoretical foundations, which involves formulating comprehensive definitions, metrics, and benchmarks.

Minimal Trade-off: It is crucial to accomplish a balanced effectiveness in the outcomes of the unified framework for addressing both shifts. Thus, future research should concentrate on developing well-crafted techniques that can achieve effective adaptation and detection without compromising one for the other in a variety of data shift scenarios.

Unified Shift Datasets: Furthermore, innovation in algorithms that can automatically adapt to different types of shifts [244], [245] is crucial, aimed at enhancing model adaptability and robustness with minimal human intervention. Equally important is the establishment of benchmark datasets and evaluation protocols that reflect real-world scenarios involving combined distribution shifts, facilitating more accurate assessments of model performance.

Interdisciplinary Approach: An interdisciplinary approach, incorporating insights from fields such as causality [246], cognitive science [247], and collaboration with domain experts, can be vital to forge direction in implementing novel frameworks to effectively develop solution to tackle data distribution shift problem.

8 Conclusion

In conclusion, this review paper has highlighted the significant obstacles posed by covariate and semantic shifts in ML and emphasized the methodologies inherent in handling these shifts independently. Also, by advocating for an integrated approach, we propose for a paradigm shift towards developing methodologies that cover the entire spectrum of distribution shifts within a single framework. This paper aims to address the current research scenario to the readers and also spark further investigation and innovation paving the way for more efficient, and effective ML applications robust to data distribution shifts.

Acknowledgments

The authors declare no conflict of interests.

REFERENCES

[1] I. H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions," *SN Computer Science*, vol. 2, no. 6, p. 420, Aug. 2021. [Online]. Available: https://doi.org/10.1007/s42979-021-00815-1

[2] J. G. Moreno-Torres, T. Raeder, R. Alaiz-Rodríguez, N. V. Chawla, and F. Herrera, "A unifying view on dataset shift in classification," *Pattern Recognition*, vol. 45, no. 1, pp. 521–530, Jan. 2012.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," Feb. 2014.

[4] J. Quiñonero-Candela, Ed., *Dataset Shift in Machine Learning*, ser. Neural Information Processing Series. Cambridge, Mass: MIT Press, 2009.

[5] A. A. Alemi, I. Fischer, and J. V. Dillon, "Uncertainty in the Variational Information Bottleneck," Jul. 2018.

[6] J. Liu, Z. Shen, Y. He, X. Zhang, R. Xu, H. Yu, and P. Cui, "Towards out-of-distribution generalization: A survey," *arXiv preprint arXiv:2108.13624*, 2021.

[7] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *arXiv preprint arXiv:2110.11334*, 2021.

[8] C. Geng, S.-j. Huang, and S. Chen, "Recent advances in open set recognition: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3614–3631, 2020.

[9] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[10] M. Pivk and F. R. Le Diberder, "Plots: A statistical tool to unfold data distributions," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 555, no. 1-2, pp. 356–369, 2005.

[11] W. Bryc, *The normal distribution: characterizations with applications*. Springer Science & Business Media, 2012, vol. 100.

[12] L. Kuipers and H. Niederreiter, *Uniform distribution of sequences*. Courier Corporation, 2012.

[13] R. B. Arellano-Valle and M. G. Genton, "On fundamental skew distributions," *Journal of Multivariate Analysis*, vol. 96, no. 1, pp. 93–116, 2005.

[14] E. A. Murphy, "One cause? many causes?: The argument from the bimodal distribution," *Journal of Chronic Diseases*, vol. 17, no. 4, pp. 301–324, 1964.

[15] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.

[16] C. Bishop, "Pattern recognition and machine learning," *Springer google schola*, vol. 2, pp. 5–43, 2006.

[17] J. S. Ramberg, E. J. Dudewicz, P. R. Tadikamalla, and E. F. Mykytka, "A probability distribution and its uses in fitting data," *Technometrics*, vol. 21, no. 2, pp. 201–214, 1979.

[18] R. Alaiz-Rodríguez, A. Guerrero-Curieses, and J. Cid-Sueiro, "Improving Classification under Changes in Class and Within-Class Distributions," in *Bio-Inspired Systems: Computational and Ambient Intelligence*, J. Cabestany, F. Sandoval, A. Prieto, and J. M. Corchado, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, vol. 5517, pp. 122–130.

[19] T. Tommasi, M. Lanzi, P. Russo, and B. Caputo, "Learning the Roots of Visual Domain Shift," in *Computer Vision – ECCV 2016 Workshops*, G. Hua and H. Jégou, Eds. Cham: Springer International Publishing, 2016, vol. 9915, pp. 475–482.

[20] C. Geng, S.-j. Huang, and S. Chen, "Recent Advances in Open Set Recognition: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3614–3631, Oct. 2021.

[21] Z. Xia, G. Dong, P. Wang, and H. Liu, "Spatial Location Constraint Prototype Loss for Open Set Recognition," Nov. 2021.

[22] V. N. Vapnik, V. Vapnik *et al.*, "Statistical learning theory," 1998.

[23] M. Sugiyama, M. Krauledat, and K.-R. Müller, "Covariate shift adaptation by importance weighted cross validation." *Journal of Machine Learning Research*, vol. 8, no. 5, 2007.

[24] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv preprint arXiv:1907.02893*, 2019.

[25] C. Zhou, X. Ma, P. Michel, and G. Neubig, "Examining and combating spurious features under distribution shift," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 857–12 867.

[26] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.

[27] A. Gesmundo and J. Dean, "An evolutionary approach to dynamic introduction of tasks in large-scale multitask learning systems," *arXiv preprint arXiv:2205.12755*, 2022.

[28] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 491–507.

[29] Y. Zhang, Q. Sun, Y. Zhou, Z. He, Z. Yin, K. Wang, L. Sheng, Y. Qiao, J. Shao, and Z. Liu, "Bamboo: Building mega-scale vision dataset continually with human-machine synergy," *arXiv preprint arXiv:2203.07845*, 2022.

[30] R. Dagli, "Astroformer: More data might not be all you need for classification," *arXiv preprint arXiv:2304.05350*, 2023.

[31] K. Yuan, S. Guo, Z. Liu, A. Zhou, F. Yu, and W. Wu, "Incorporating convolution designs into visual transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 579–588.

[32] Y. Huang, Y. Cheng, A. Bapna, O. Firat, D. Chen, M. Chen, H. Lee, J. Ngiam, Q. V. Le, Y. Wu *et al.*, "Gpipe: Efficient training of giant neural networks using pipeline parallelism," *Advances in neural information processing systems*, vol. 32, 2019.

[33] V. Macko, C. Weill, H. Mazzawi, and J. Gonzalvo, "Improving neural architecture search image classifiers via ensemble learning," *arXiv preprint arXiv:1903.06236*, 2019.

[34] M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Conditional adversarial domain adaptation," *Advances in neural information processing systems*, vol. 31, 2018.

[35] T. Westfechtel, D. Zhang, and T. Harada, "Combining inherent knowledge of vision-language models with unsupervised domain adaptation through self-knowledge distillation," *CoRR*, 2023.

[36] V. H. Phan, T. L. Tran, Q. Tran, and T. Le, "Enhancing domain adaptation through prompt gradient alignment," *Advances in Neural Information Processing Systems*, vol. 37, pp. 45 518–45 551, 2024.

[37] J. Li, E. Chen, Z. Ding, L. Zhu, K. Lu, and H. T. Shen, "Maximum density divergence for domain adaptation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3918–3930, 2020.

[38] S. Cui, S. Wang, J. Zhuo, C. Su, Q. Huang, and Q. Tian, "Gradually vanishing bridge for adversarial domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 455–12 464.

[39] J. Na, H. Jung, H. J. Chang, and W. Hwang, "Fixbi: Bridging domain spaces for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1094–1103.

[40] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *International conference on machine learning*. PMLR, 2020, pp. 6028–6039.

[41] Y. Yang and S. Soatto, "Fda: Fourier domain adaptation for semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4085–4095.

[42] T. Westfechtel, H.-W. Yeh, D. Zhang, and T. Harada, "Gradual source domain expansion for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 1946–1955.

[43] A. Saporta, T.-H. Vu, M. Cord, and P. Pérez, "Esl: Entropy-guided self-supervised learning for domain adaptation in semantic segmentation," *arXiv preprint arXiv:2006.08658*, 2020.

[44] I. Gulrajani and D. Lopez-Paz, "In search of lost domain generalization," *arXiv preprint arXiv:2007.01434*, 2020.

[45] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[46] H. Li, S. J. Pan, S. Wang, and A. C. Kot, "Domain generalization with adversarial feature learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5400–5409.

[47] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[48] J. Cha, S. Chun, K. Lee, H.-C. Cho, S. Park, Y. Lee, and S. Park, "Swad: Domain generalization by seeking flat minima," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 405–22 418, 2021.

[49] J. Cho, G. Nam, S. Kim, H. Yang, and S. Kwak, "Promptstyler: Prompt-driven style generation for source-free domain generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 702–15 712.

[50] Z. Li, K. Ren, X. Jiang, Y. Shen, H. Zhang, and D. Li, "Simple: Specialized model-sample matching for domain generalization," in *The Eleventh International Conference on Learning Representations*, 2022.

[51] S. Bai, Y. Zhang, W. Zhou, Z. Luan, and B. Chen, "Soft prompt generation for domain generalization," in *European Conference on Computer Vision*. Springer, 2024, pp. 434–450.

[52] Z. Tan, X. Yang, and K. Huang, "Rethinking multi-domain generalization with a general learning objective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 512–23 522.

[53] L. Ruff, R. A. Vandermeulen, N. Görnitz, A. Binder, E. Müller, K.-R. Müller, and M. Kloft, "Deep semi-supervised anomaly detection," *arXiv preprint arXiv:1906.02694*, 2019.

[54] F. Rosenblatt, "The perceptron: a probabilistic model for information storage and organization in the brain." *Psychological review*, vol. 65, no. 6, p. 386, 1958.

[55] G. Pang, C. Shen, H. Jin, and A. van den Hengel, "Deep weakly-supervised anomaly detection," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 1795–1807.

[56] K.-L. Li, H.-K. Huang, S.-F. Tian, and W. Xu, "Improving one-class svm for anomaly detection," in *Proceedings of the 2003 international conference on machine learning and cybernetics (IEEE Cat. No. 03EX693)*, vol. 5. IEEE, 2003, pp. 3077–3081.

[57] G. Pang, L. Cao, L. Chen, and H. Liu, "Learning representations of ultrahigh-dimensional data for random distance-based outlier detection," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 2041–2050.

[58] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[59] G. Pang, C. Shen, and A. Van Den Hengel, "Deep anomaly detection with deviation networks," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 353–362.

[60] Y. Zhao and M. K. Hryniewicki, "Xgbod: improving supervised outlier detection with unsupervised representation learning," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[61] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems*, vol. 30, 2017.

[62] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[63] A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572.

[64] Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative openmax for multi-class open set classification," *arXiv preprint arXiv:1707.07418*, 2017.

[65] D.-W. Zhou, H.-J. Ye, and D.-C. Zhan, "Learning placeholders for open-set recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4401–4410.

[66] P. Oza and V. M. Patel, "C2ae: Class conditioned auto-encoder for open-set recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2307–2316.

[67] S. Kong and D. Ramanan, "Opengan: Open-set recognition via open data generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 813–822.

[68] G. Chen, P. Peng, X. Wang, and Y. Tian, "Adversarial reciprocal points learning for open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8065–8081, 2021.

[69] D. Miller, N. Sunderhauf, M. Milford, and F. Dayoub, "Class anchor clustering: A loss for distance-based open set recognition," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3570–3578.

[70] H. Huang, Y. Wang, Q. Hu, and M.-M. Cheng, "Class-Specific Semantic Reconstruction for Open Set Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–14, 2022.

[71] D. Hendrycks, A. Zou, M. Mazeika, L. Tang, B. Li, D. Song, and J. Steinhardt, "Pixmix: Dreamlike pictures comprehensively improve safety measures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 783–16 792.

[72] H. Wei, R. Xie, H. Cheng, L. Feng, B. An, and Y. Li, "Mitigating neural network overconfidence with logit normalization," in *International conference on machine learning*. PMLR, 2022, pp. 23 631–23 644.

[73] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," *arXiv preprint arXiv:1812.04606*, 2018.

[74] Y. Zou, W. W. Ng, X. Zhang, B. Loo, X. Yan, and R. Wang, "Mcd: Defense against query-based black-box surrogate attacks," in *2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2024, pp. 5359–5366.

[75] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song, "Using self-supervised learning can improve model robustness and uncertainty," *Advances in neural information processing systems*, vol. 32, 2019.

[76] J. Tack, S. Mo, J. Jeong, and J. Shin, "Csi: Novelty detection via contrastive learning on distributionally shifted instances," *Advances in neural information processing systems*, vol. 33, pp. 11 839–11 852, 2020.

[77] J. Zhang, N. Inkawhich, R. Linderman, Y. Chen, and H. Li, "Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5531–5540.

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

[79] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, "Experience replay for continual learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[80] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[81] Z. Wang, L. Liu, Y. Kong, J. Guo, and D. Tao, "Online continual learning with contrastive vision transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 631–650.

[82] F. Sarfraz, E. Arani, and B. Zonooz, "Sparse coding in a dual memory system for lifelong learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 8, 2023, pp. 9714–9722.

[83] Q. Pham, C. Liu, and S. Hoi, "Dualnet: Continual learning, fast and slow," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 131–16 144, 2021.

[84] X. Nie, S. Xu, X. Liu, G. Meng, C. Huo, and S. Xiang, "Bilateral memory consolidation for continual learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 16 026–16 035.

[85] Z. Man, K. Huang, Y. Zhang, Y. Chen, Y. Chen, and J. Xu, "Icl: Iterative continual learning for multi-domain neural machine translation," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 7732–7743.

[86] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, "Dark experience for general continual learning: a strong, simple baseline," *Advances in neural information processing systems*, vol. 33, pp. 15 920–15 930, 2020.

[87] H. Cha, J. Lee, and J. Shin, "Co2l: Contrastive continual learning," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 9516–9525.

[88] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," 2020.
[89] X. Wang, T.-K. Huang, and J. Schneider, "Active transfer learning under model shift," in *International Conference on Machine Learning*. PMLR, 2014, pp. 1305–1313.
[90] T. Joachims *et al.*, "Transductive inference for text classification using support vector machines," in *Icml*, vol. 99, 1999, pp. 200–209.
[91] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.
[92] C. B. Do and A. Y. Ng, "Transfer learning for text classification," *Advances in neural information processing systems*, vol. 18, 2005.
[93] R. Raina, A. Y. Ng, and D. Koller, "Constructing informative priors using transfer learning," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 713–720.
[94] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3712–3722.
[95] M. Long, H. Zhu, J. Wang, and M. I. Jordan, "Deep transfer learning with joint adaptation networks," in *International conference on machine learning*. PMLR, 2017, pp. 2208–2217.
[96] B. Neyshabur, H. Sedghi, and C. Zhang, "What is being transferred in transfer learning?" *Advances in neural information processing systems*, vol. 33, pp. 512–523, 2020.
[97] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," in *Proceedings of ICML workshop on unsupervised and transfer learning*. JMLR Workshop and Conference Proceedings, 2012, pp. 17–36.
[98] B. Cao, S. J. Pan, Y. Zhang, D.-Y. Yeung, and Q. Yang, "Adaptive transfer learning," in *proceedings of the AAAI Conference on Artificial Intelligence*, vol. 24, no. 1, 2010, pp. 407–412.
[99] P. Zhao, S. C. Hoi, J. Wang, and B. Li, "Online transfer learning," *Artificial intelligence*, vol. 216, pp. 76–102, 2014.
[100] X. Wang and J. Schneider, "Flexible transfer learning under support and model shift," *Advances in Neural Information Processing Systems*, vol. 27, 2014.
[101] X. Liu, Y. Li, Q. Meng, and G. Chen, "Deep transfer learning for conditional shift in regression," *Knowledge-Based Systems*, vol. 227, p. 107216, 2021.
[102] G. Chen, Y. Li, and X. Liu, "Transfer learning under conditional shift based on fuzzy residual," *IEEE Transactions on Cybernetics*, vol. 52, no. 2, pp. 960–970, 2020.
[103] Z. Zhu, K. Lin, A. K. Jain, and J. Zhou, "Transfer learning in deep reinforcement learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[104] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC medical imaging*, vol. 22, no. 1, p. 69, 2022.
[105] W. Li, R. Huang, J. Li, Y. Liao, Z. Chen, G. He, R. Yan, and K. Gryllias, "A perspective survey on deep transfer learning for fault diagnosis in industrial scenarios: Theories, applications and challenges," *Mechanical Systems and Signal Processing*, vol. 167, p. 108487, 2022.
[106] J. Y.-L. Chan, K. T. Bea, S. M. H. Leow, S. W. Phoong, and W. K. Cheng, "State of the art: a review of sentiment analysis based on sequential transfer learning," *Artificial Intelligence Review*, vol. 56, no. 1, pp. 749–780, 2023.
[107] S. T. Mehedi, A. Anwar, Z. Rahman, K. Ahmed, and R. Islam, "Dependable intrusion detection system for iot: A deep transfer learning based approach," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 1006–1017, 2022.
[108] M. Wang and W. Deng, "Deep visual domain adaptation: A survey," *Neurocomputing*, vol. 312, pp. 135–153, 2018.
[109] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.
[110] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto, "Unified deep supervised domain adaptation and generalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 5715–5725.
[111] A. Saha, P. Rai, H. Daumé, S. Venkatasubramanian, and S. L. DuVall, "Active supervised domain adaptation," in *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2011, Athens, Greece, September 5-9, 2011, Proceedings, Part III 22*. Springer, 2011, pp. 97–112.
[112] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5058–5062.
[113] P. Koniusz, Y. Tas, H. Zhang, M. Harandi, F. Porikli, and R. Zhang, "Museum exhibit identification challenge for the supervised domain adaptation and beyond," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 788–804.
[114] K. Saito, D. Kim, S. Sclaroff, T. Darrell, and K. Saenko, "Semi-supervised domain adaptation via minimax entropy," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8050–8058.
[115] H. Daumé III, A. Kumar, and A. Saha, "Frustratingly easy semi-supervised domain adaptation," in *Proceedings of the 2010 Workshop on Domain Adaptation for Natural Language Processing*, 2010, pp. 53–59.
[116] G. He, X. Liu, F. Fan, and J. You, "Classification-aware semi-supervised domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 964–965.
[117] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.
[118] G. Kang, L. Jiang, Y. Yang, and A. G. Hauptmann, "Contrastive adaptation network for unsupervised domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4893–4902.
[119] R. Li, Q. Jiao, W. Cao, H.-S. Wong, and S. Wu, "Model adaptation: Unsupervised domain adaptation without source data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9641–9650.
[120] P. Panareda Busto and J. Gall, "Open set domain adaptation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 754–763.
[121] Z. Cao, M. Long, J. Wang, and M. I. Jordan, "Partial transfer learning with selective adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2724–2732.
[122] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Universal domain adaptation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2720–2729.
[123] M. Wulfmeier, A. Bewley, and I. Posner, "Incremental adversarial domain adaptation for continually changing environments," in *2018 IEEE International conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 4489–4495.
[124] K. Zhang, B. Schölkopf, K. Muandet, and Z. Wang, "Domain adaptation under target and conditional shift," in *International conference on machine learning*. PMLR, 2013, pp. 819–827.
[125] S. Garg, S. Balakrishnan, and Z. Lipton, "Domain adaptation under open set label shift," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 531–22 546, 2022.
[126] S. Garg, N. Erickson, J. Sharpnack, A. Smola, S. Balakrishnan, and Z. C. Lipton, "Rlsbench: Domain adaptation under relaxed label shift," in *International Conference on Machine Learning*. PMLR, 2023, pp. 10 879–10 928.
[127] F. Qiao, L. Zhao, and X. Peng, "Learning to learn single domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 556–12 565.
[128] H. Ye, C. Xie, T. Cai, R. Li, Z. Li, and L. Wang, "Towards a theoretical framework of out-of-distribution generalization," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 519–23 531, 2021.
[129] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville, "Out-of-distribution generalization via risk extrapolation (rex)," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5815–5826.
[130] M. W. Spratling, "Comprehensive assessment of the performance of deep learning classifiers reveals a surprising lack of robustness," *arXiv preprint arXiv:2308.04137*, 2023.
[131] C.-Y. Chuang, A. Torralba, and S. Jegelka, "Estimating generalization under distribution shifts via domain-invariant representations," *arXiv preprint arXiv:2007.03511*, 2020.
[132] M. Ding, K. Kong, J. Chen, J. Kirchenbauer, M. Goldblum, D. Wipf, F. Huang, and T. Goldstein, "A closer look at

distribution shifts and out-of-distribution generalization on graphs," 2021.
[133] Y. Chen, Y. Zhang, Y. Bian, H. Yang, M. Kaili, B. Xie, T. Liu, B. Han, and J. Cheng, "Learning causally invariant representations for out-of-distribution generalization on graphs," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22 131–22 148, 2022.
[134] K. Sohn, H. Chang, J. Lezama, L. Polania, H. Zhang, Y. Hao, I. Essa, and L. Jiang, "Visual prompt tuning for generative transfer learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 840–19 851.
[135] Q. Qian, Y. Qin, J. Luo, Y. Wang, and F. Wu, "Deep discriminative transfer learning network for cross-machine fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 186, p. 109884, 2023.
[136] R. Zhu, W. Peng, D. Wang, and C.-G. Huang, "Bayesian transfer learning with active querying for intelligent cross-machine fault prognosis under limited data," *Mechanical Systems and Signal Processing*, vol. 183, p. 109628, 2023.
[137] D. A. Bierbrauer, M. J. De Lucia, K. Reddy, P. Maxwell, and N. D. Bastian, "Transfer learning for raw network traffic detection," *Expert Systems with Applications*, vol. 211, p. 118641, 2023.
[138] X. Zhou, Q. Yang, Q. Liu, W. Liang, K. Wang, Z. Liu, J. Ma, and Q. Jin, "Spatial–temporal federated transfer learning with multi-sensor data fusion for cooperative positioning," *Information Fusion*, vol. 105, p. 102182, 2024.
[139] Q. Xiao, Y. Zhang, and Q. Yang, "Selective random walk for transfer learning in heterogeneous label spaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
[140] Z. Li, R. Cai, G. Chen, B. Sun, Z. Hao, and K. Zhang, "Subspace identification for multi-source domain adaptation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[141] L. Hoyer, D. Dai, H. Wang, and L. Van Gool, "Mic: Masked image consistency for context-enhanced domain adaptation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 721–11 732.
[142] T.-D. Truong, N. Le, B. Raj, J. Cothren, and K. Luu, "Fredom: Fairness domain adaptation approach to semantic scene understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 988–19 997.
[143] G. Kim and S. Y. Chun, "Datid-3d: Diversity-preserved domain adaptation using text-to-image diffusion for 3d generative model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 203–14 213.
[144] S. Wang, B. Wang, Z. Zhang, A. A. Heidari, and H. Chen, "Class-aware sample reweighting optimal transport for multi-source domain adaptation," *Neurocomputing*, vol. 523, pp. 213–223, 2023.
[145] Y. Cheng, F. Wei, J. Bao, D. Chen, and W. Zhang, "Adpl: Adaptive dual path learning for domain adaptation of semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
[146] X. Hao, W. Zhang, D. Wu, F. Zhu, and B. Li, "Dual alignment unsupervised domain adaptation for video-text retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 962–18 972.
[147] C. Ge, R. Huang, M. Xie, Z. Lai, S. Song, S. Li, and G. Huang, "Domain adaptation via prompt learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
[148] P. Wang, Z. Zhang, Z. Lei, and L. Zhang, "Sharpness-aware gradient matching for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3769–3778.
[149] M. Segu, A. Tonioni, and F. Tombari, "Batch normalization embeddings for deep domain generalization," *Pattern Recognition*, vol. 135, p. 109115, 2023.
[150] J. Chen, M. Jiang, Q. Dou, and Q. Chen, "Federated domain generalization for image recognition via cross-client style transfer," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 361–370.
[151] J. Yu, J. Liang, and R. He, "Mind the label shift of augmentation-based graph ood generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 620–11 630.
[152] M. Zhang, J. Yuan, Y. He, W. Li, Z. Chen, and K. Kuang, "Map: Towards balanced generalization of iid and ood through model-agnostic adapters," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 11 921–11 931.
[153] H. Wang, K. Kuang, L. Lan, Z. Wang, W. Huang, F. Wu, and W. Yang, "Out-of-distribution generalization with causal feature separation," *IEEE Transactions on Knowledge and Data Engineering*, 2023.
[154] T. M. Mitchell, *Machine Learning*, nachdr. ed., ser. McGraw-Hill Series in Computer Science. New York: McGraw-Hill, 2013.
[155] A. Tsymbal, "The problem of concept drift: definitions and related work," *Computer Science Department, Trinity College Dublin*, vol. 106, no. 2, p. 58, 2004.
[156] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 951–10 960.
[157] J. Tian, Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Exploring covariate and concept shift for detection and confidence calibration of out-of-distribution data," 2021.
[158] J. Parmar, S. Chouhan, V. Raychoudhury, and S. Rathore, "Open-world machine learning: applications, challenges, and opportunities," *ACM Computing Surveys*, vol. 55, no. 10, pp. 1–37, 2023.
[159] A. Mahdavi and M. Carvalho, "A Survey on Open Set Recognition," in *2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Dec. 2021, pp. 37–44.
[160] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.
[161] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.
[162] H. Zhang and V. M. Patel, "Sparse representation-based open set recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 8, pp. 1690–1696, 2016.
[163] L. Tamang, M. R. Bouadjenek, R. Dazeley, and S. Aryal, "Margin-bounded confidence scores for out-of-distribution detection," in *2024 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2024, pp. 1–10.
[164] Z. Zhang and X. Xiang, "Decoupling maxlogit for out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3388–3397.
[165] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," *Advances in Neural Information Processing Systems*, vol. 34, pp. 144–157, 2021.
[166] R. Huang, A. Geng, and Y. Li, "On the importance of gradients for detecting distributional shifts in the wild," *Advances in Neural Information Processing Systems*, vol. 34, pp. 677–689, 2021.
[167] X. Wu, J. Lu, Z. Fang, and G. Zhang, "Meta ood learning for continuously adaptive ood detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 353–19 364.
[168] P. Morteza and Y. Li, "Provable guarantees for understanding out-of-distribution detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, 2022, pp. 7831–7840.
[169] Z. Fang, Y. Li, J. Lu, J. Dong, B. Han, and F. Liu, "Is out-of-distribution detection learnable?" *Advances in Neural Information Processing Systems*, vol. 35, pp. 37 199–37 213, 2022.
[170] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in neural information processing systems*, vol. 32, 2019.
[171] S. Mohseni, M. Pitale, J. Yadawa, and Z. Wang, "Self-supervised learning for generalizable out-of-distribution detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, 2020, pp. 5216–5223.
[172] F. Zhu, Z. Cheng, X.-Y. Zhang, and C.-L. Liu, "Openmix: Exploring outlier samples for misclassification detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 074–12 083.

[173] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *Advances in neural information processing systems*, vol. 33, pp. 21 464–21 475, 2020.
[174] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha, "Atom: Robustifying out-of-distribution detection using outlier mining," in *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*. Springer, 2021, pp. 430–445.
[175] Y. Ming, Y. Fan, and Y. Li, "Poem: Out-of-distribution detection with posterior sampling," in *International Conference on Machine Learning*. PMLR, 2022, pp. 15 650–15 665.
[176] J. Katz-Samuels, J. B. Nakhleh, R. Nowak, and Y. Li, "Training ood detectors in their natural habitats," in *International Conference on Machine Learning*. PMLR, 2022, pp. 10 848–10 865.
[177] M. Masana, I. Ruiz, J. Serrat, J. van de Weijer, and A. M. Lopez, "Metric learning for novelty and anomaly detection," *arXiv preprint arXiv:1808.05492*, 2018.
[178] E. R. Faria, I. J. Gonçalves, A. C. de Carvalho, and J. Gama, "Novelty detection in data streams," *Artificial Intelligence Review*, vol. 45, pp. 235–269, 2016.
[179] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli, "Adversarially learned one-class classifier for novelty detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3379–3388.
[180] D. Abati, A. Porrello, S. Calderara, and R. Cucchiara, "Latent space autoregression for novelty detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 481–490.
[181] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 8765–8775, 2020.
[182] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
[183] H. Cho, J. Seol, and S.-g. Lee, "Masked contrastive learning for anomaly detection," *arXiv preprint arXiv:2105.08793*, 2021.
[184] O. Kopuklu, J. Zheng, H. Xu, and G. Rigoll, "Driver anomaly detection: A dataset and contrastive learning approach," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 91–100.
[185] R. Hadsell, D. Rao, A. A. Rusu, and R. Pascanu, "Embracing change: Continual learning in deep neural networks," *Trends in cognitive sciences*, vol. 24, no. 12, pp. 1028–1040, 2020.
[186] T. Lesort, M. Caccia, and I. Rish, "Understanding continual learning settings with data distribution drift analysis," *arXiv preprint arXiv:2104.01678*, 2021.
[187] G. Fischer, "Lifelong learning—more than training," *Journal of Interactive Learning Research*, vol. 11, no. 3, pp. 265–294, 2000.
[188] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 233–248.
[189] R. Aljundi, K. Kelchtermans, and T. Tuytelaars, "Task-free continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 254–11 263.
[190] J. Schwarz, W. Czarnecki, J. Luketina, A. Grabska-Barwinska, Y. W. Teh, R. Pascanu, and R. Hadsell, "Progress & compress: A scalable framework for continual learning," in *International conference on machine learning*. PMLR, 2018, pp. 4528–4537.
[191] A. M. N. Taufique, C. S. Jahan, and A. Savakis, "Unsupervised continual learning for gradually varying domains," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3740–3750.
[192] L. Caccia, R. Aljundi, N. Asadi, T. Tuytelaars, J. Pineau, and E. Belilovsky, "New insights on reducing abrupt representation change in online continual learning," *arXiv preprint arXiv:2104.05025*, 2021.
[193] M. Bashari, A. Epstein, Y. Romano, and M. Sesia, "Derandomized novelty detection with fdr control via conformal e-values," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
[194] F. Zhu, W. Zhang, X. Chen, X. Gao, and N. Ye, "Large margin distribution multi-class supervised novelty detection," *Expert Systems with Applications*, vol. 224, p. 119937, 2023.
[195] W. Liu, H. Chang, B. Ma, S. Shan, and X. Chen, "Diversity-measurable anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 12 147–12 156.
[196] H. Xu, Z. Sun, Y. Cao, and H. Bilal, "A data-driven approach for intrusion and anomaly detection using automated machine learning for the internet of things," *Soft Computing*, vol. 27, no. 19, pp. 14 469–14 481, 2023.
[197] Y. Chen, Z. Liu, B. Zhang, W. Fok, X. Qi, and Y.-C. Wu, "Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 387–395.
[198] G. Xie, J. Wang, J. Liu, F. Zheng, and Y. Jin, "Pushing the limits of fewshot anomaly detection in industry vision: Graphcore," *arXiv preprint arXiv:2301.12082*, 2023.
[199] J. Wang, X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, H. Huang, W. Ye, X. Geng *et al.*, "On the robustness of chatgpt: An adversarial and out-of-distribution perspective," *arXiv preprint arXiv:2302.12095*, 2023.
[200] M. S. Graham, W. H. Pinaya, P.-D. Tudosiu, P. Nachev, S. Ourselin, and J. Cardoso, "Denoising diffusion models for out-of-distribution detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2947–2956.
[201] Q. Wu, Y. Chen, C. Yang, and J. Yan, "Energy-based out-of-distribution detection for graph neural networks," *arXiv preprint arXiv:2302.02914*, 2023.
[202] S. Wilson, T. Fischer, N. Sünderhauf, and F. Dayoub, "Hyperdimensional feature fusion for out-of-distribution detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2644–2654.
[203] D. Song, Z. Wang, Y. Huang, L. Ma, and T. Zhang, "Deeplens: Interactive out-of-distribution data detection in nlp models," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–17.
[204] J. Li, P. Chen, Z. He, S. Yu, S. Liu, and J. Jia, "Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 11 578–11 589.
[205] J. Mereau and A. Minaro, "Detecting textual out-of-distribution examples with pretrained transformers."
[206] C. Liu, C. Yang, H.-B. Qin, X. Zhu, C.-L. Liu, and X.-C. Yin, "Towards open-set text recognition via label-to-prototype learning," *Pattern Recognition*, vol. 134, p. 109109, 2023.
[207] J. Liu, J. Tian, W. Han, Z. Qin, Y. Fan, and J. Shao, "Learning multiple gaussian prototypes for open-set recognition," *Information Sciences*, vol. 626, pp. 738–753, 2023.
[208] J. Sun, H. Wang, and Q. Dong, "Hierarchical attention network for open-set fine-grained image recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
[209] T. Yang, D. Wang, F. Tang, X. Zhao, J. Cao, and S. Tang, "Progressive open space expansion for open-set model attribution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 856–15 865.
[210] H. Zhang, Y. Liu, Y. Wang, L. Wang, and Y. Qiao, "Learning discriminative feature representation for open set action recognition," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7696–7705.
[211] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee, "Gligen: Open-set grounded text-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 511–22 521.
[212] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.
[213] M. Soltani, B. Ousat, M. J. Siavoshani, and A. H. Jahangir, "An adaptable deep learning-based intrusion detection system to zero-day attacks," *Journal of Information Security and Applications*, vol. 76, p. 103516, 2023.
[214] J. S. Smith, J. Tian, S. Halbe, Y.-C. Hsu, and Z. Kira, "A closer look at rehearsal-free continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2409–2419.

[215] A. Villa, J. L. Alcázar, M. Alfarra, K. Alhamoud, J. Hurtado, F. C. Heilbron, A. Soto, and B. Ghanem, "Pivot: Prompting for video continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 214–24 223.
[216] A. Razdaibiedina, Y. Mao, R. Hou, M. Khabsa, M. Lewis, and A. Almahairi, "Progressive prompts: Continual learning for language models," *arXiv preprint arXiv:2301.12314*, 2023.
[217] L. Yuan, Y. Ma, L. Su, and Z. Wang, "Peer-to-peer federated continual learning for naturalistic driving action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5249–5258.
[218] Y. Yang, Z. Cui, J. Xu, C. Zhong, W.-S. Zheng, and R. Wang, "Continual learning with bayesian model based on a fixed pre-trained feature extractor," *Visual Intelligence*, vol. 1, no. 1, p. 5, 2023.
[219] L. Zhu, T. Chen, J. Yin, S. See, and J. Liu, "Continual semantic segmentation with automatic memory sample selection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3082–3092.
[220] U. Topcu, N. Bliss, N. Cooke, M. Cummings, A. Llorens, H. Shrobe, and L. Zuck, "Assured autonomy: Path toward living with autonomous systems we can trust," *arXiv preprint arXiv:2010.14443*, 2020.
[221] V. Hashemi, J. Křetínskỳ, S. Rieder, and J. Schmidt, "Runtime monitoring for out-of-distribution detection in object detection neural networks," in *International Symposium on Formal Methods*. Springer, 2023, pp. 622–634.
[222] J. Guérin, K. Delmas, R. Ferreira, and J. Guiochet, "Out-of-distribution detection is not all you need," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 37, no. 12, 2023, pp. 14 829–14 837.
[223] A. Roy, A. Cobb, N. Bastian, B. Jalaian, and S. Jha, "Runtime monitoring of deep neural networks using top-down context models inspired by predictive processing and dual process theory," in *AAAI Spring Symposium 2022*, 2022.
[224] A. Bendale and T. Boult, "Towards open world recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1893–1902.
[225] T. E. Boult, S. Cruz, A. R. Dhamija, M. Gunther, J. Henrydoss, and W. J. Scheirer, "Learning and the unknown: Surveying steps toward open world recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9801–9807.
[226] B. Romera-Paredes and P. Torr, "An embarrassingly simple approach to zero-shot learning," in *International conference on machine learning*. PMLR, 2015, pp. 2152–2161.
[227] S. Min, H. Yao, H. Xie, C. Wang, Z.-J. Zha, and Y. Zhang, "Domain-aware visual bias eliminating for generalized zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 664–12 673.
[228] J. Wu, T. Zhang, Z.-J. Zha, J. Luo, Y. Zhang, and F. Wu, "Self-supervised domain-aware generative network for generalized zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 767–12 776.
[229] U. Maniyar, A. A. Deshmukh, U. Dogan, V. N. Balasubramanian *et al.*, "Zero shot domain generalization," *arXiv preprint arXiv:2008.07443*, 2020.
[230] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4051–4070, 2022.
[231] W. Wang, V. W. Zheng, H. Yu, and C. Miao, "A survey of zero-shot learning: Settings, methods, and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–37, 2019.
[232] L. Yang, Y. Song, X. Ren, C. Lyu, Y. Wang, J. Zhuo, L. Liu, J. Wang, J. Foster, and Y. Zhang, "Out-of-distribution generalization in natural language processing: Past, present, and future," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 4533–4559.
[233] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.
[234] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International conference on machine learning*. PMLR, 2021, pp. 5637–5664.
[235] T. Wang, A. Roberts, D. Hesslow, T. Le Scao, H. W. Chung, I. Beltagy, J. Launay, and C. Raffel, "What language model architecture and pretraining objective works best for zero-shot generalization?" in *International Conference on Machine Learning*. PMLR, 2022, pp. 22 964–22 984.
[236] K. S. Desale, *Concept Drift in Large Language Models: Adapting the Conversation*. CRC Press, 2025.
[237] L. Yuan, Y. Chen, G. Cui, H. Gao, F. Zou, X. Cheng, H. Ji, Z. Liu, and M. Sun, "Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations," *Advances in Neural Information Processing Systems*, vol. 36, pp. 58 478–58 507, 2023.
[238] A. Filos, P. Tigkas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal, "Can autonomous vehicles identify, recover from, and adapt to distribution shifts?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 3145–3153.
[239] Z. Yang, X. He, J. Zhang, J. Wu, X. Xin, J. Chen, and X. Wang, "A generic learning framework for sequential recommendation with distribution shifts," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023, pp. 331–340.
[240] X. Wang, Q. Kang, J. An, and M. Zhou, "Drifted twitter spam classification using multiscale detection test on kl divergence," *IEEE Access*, vol. 7, pp. 108 384–108 394, 2019.
[241] H. Bai, G. Canal, X. Du, J. Kwon, R. D. Nowak, and Y. Li, "Feed two birds with one scone: Exploiting wild data for both out-of-distribution generalization and detection," in *International Conference on Machine Learning*. PMLR, 2023, pp. 1454–1471.
[242] J. Yang, K. Zhou, and Z. Liu, "Full-spectrum out-of-distribution detection," *International Journal of Computer Vision*, pp. 1–16, 2023.
[243] X. Zhang, Y. He, R. Xu, H. Yu, Z. Shen, and P. Cui, "Nico++: Towards better benchmarking for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 036–16 047.
[244] Z. Gao, B. Li, M. Salzmann, and X. He, "Generalize or detect? towards robust semantic segmentation under multiple distribution shifts," *Advances in Neural Information Processing Systems*, vol. 37, pp. 52 014–52 039, 2024.
[245] J.-J. Shao, X.-W. Yang, and L.-Z. Guo, "Open-set learning under covariate shift," *Machine Learning*, vol. 113, no. 4, pp. 1643–1659, 2024.
[246] F. Lv, J. Liang, S. Li, B. Zang, C. H. Liu, Z. Wang, and D. Liu, "Causality inspired representation learning for domain generalization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8046–8056.
[247] M. H. Bickhard and L. Terveen, *Foundational issues in artificial intelligence and cognitive science: Impasse and solution*. Elsevier, 1995, vol. 109.