From Local Geometry to Global Pseudo-Labeling for Robust Positive–Unlabeled Learning under Covariate Shift

Firas Gabetni[1,2], Alexandre Rocchi–Henry[1,2], Nacim Belkhir[1], Ziyi Liu[1], and Gianni Franchi[2]

[1] U2IS, ENSTA, Institut Polytechnique de Paris
[2] AMIAD, Pôle Recherche, Palaiseau

Abstract. Detecting covariate shift is critical for building reliable vision systems. While most prior work focuses on improving robustness to shift, explicitly detecting covariate shift remains underexplored. Existing approaches typically rely on fully supervised training, requiring labeled examples from both original and shifted distributions, which is often impractical. In this paper, we show that covariate shift detection can be effectively addressed with weaker supervision using Positive–Unlabeled (PU) learning. However, under covariate shift, in-distribution and shifted data overlap significantly, making classical PU methods unstable and sensitive to noise. To overcome this challenge, we introduce **Spectral PU Neighborhood Annotation (S-PUNA)**, a geometry-aware framework that progressively discovers shifted data by leveraging the local manifold structure of visual features. Extensive experiments show that **S-PUNA** achieves state-of-the-art performance in PU settings and remarkably matches the performance of fully supervised methods. Moreover, our approach transfers robustly across different types of shifts, demonstrating strong generalization capabilities. Code is available at `https://github.com/fira7s/S-PUNA`.

Keywords: Covariate shift · Positive Unlabeled Learning · Pseudo labeling

1 Introduction

Modern Deep Neural Networks (DNNs) are usually trained on a source distribution and tested on data assumed to follow a similar distribution. A central objective of classical machine learning research has been to improve generalization under mild distributional shifts, commonly referred to as *covariate shift*, where the input distribution P_X changes but $P(Y \mid X)$ stays the same. Many works therefore focus on domain generalization [42], domain adaptation [1], and robustness to small input perturbations [18]. In contrast, under a *semantic shift* [38] (for example when a model trained on animals is tested on objects) the predictions become unreliable. In such cases, the goal is not to generalize but to detect

the shift and raise an alert. Thus, models should be invariant to covariate shift but sensitive to semantic shift.

In this work, we argue that *detecting covariate shift itself* is also an important problem that has been less studied. In many real applications, knowing that the data distribution has changed is critical. For example, detecting AI-generated data often relies on identifying small distribution differences [39, 43]. Similarly, in finance [11], healthcare [15], or sensor systems [27], gradual covariate shifts can signal that model predictions may become unreliable.

Existing approaches are predominantly fully supervised. For example, recent work such as DisCoPatch [6] trains a dedicated discriminator to distinguish between real and synthetic images. While effective, such methods require labeled examples from both distributions, limiting their applicability when negative samples are unavailable or expensive to curate. This raises a fundamental question:

Are we forced to rely on fully supervised training for covariate shift detection, or can we relax the level of supervision?

We propose to address covariate shift detection through the lens of *Positive–Unlabeled (PU) learning*. This field studies binary classification when only labeled positive examples and unlabeled data are available. This paradigm naturally arises in anomaly detection [35], medical diagnosis [3], or out-of-distribution detection [20], where negative labels are scarce or unreliable. Surprisingly, to the best of our knowledge, PU learning has not been systematically investigated for covariate shift detection.

However, directly applying classical PU risk minimization in the presence of covariate shift is problematic. Standard PU methods rely on unbiased risk estimators that assume sufficient separability between positive and negative distributions. Under covariate shift, the distributions may overlap significantly, leading to a small total variation distance and high intrinsic Bayes risk. In this regime, global risk estimation becomes unstable and prone to high variance.

We argue that in covariate-shifted PU settings, the negative distribution should not be estimated globally. Instead, it should be *progressively discovered* through its local geometric structure. Motivated by the manifold hypothesis, we treat the unlabeled dataset as a mixture of low-dimensional submanifolds corresponding to positive and shifted negative components. Rather than directly estimating risk differences, we iteratively uncover the negative manifold by propagating reliable local neighborhood information from positive anchor points.

To operationalize this idea, we introduce **Spectral PU Neighborhood Annotation (S-PUNA)**, a geometry-aware pseudo-labeling framework for covariate shift detection as illustrated in Figure 1. **S-PUNA** initializes from labeled positives and iteratively expands both positive and negative anchor sets using confidence-controlled k-nearest neighbor propagation in feature space. Unlike classical k-NN-PU [40] that propagates only positive labels, **S-PUNA** symmetrically constructs both classes, progressively refining the decision boundary.

A key challenge in iterative pseudo-labeling is preventing semantic drift. We address this through a novel *spectral entropy stopping criterion*. By monitoring

the eigenvalue distribution of the covariance matrix of the discovered negative set, we measure its intrinsic dimensionality and structural complexity. The iterative expansion stops once the spectral entropy stabilizes, indicating that the negative manifold has been sufficiently captured without invading overlapping regions.

Finally, we train a DNN classifier to distill the non-parametric neighborhood structure to maximize latent separation and distinguish positive from shifted components.

Our contributions are summarized as follows:

- We are the first to use PU for detecting changes in covariates and to provide theoretical insights on this new problem.
- We introduce **S-PUNA**, a geometry-aware framework that progressively uncovers the shifted distribution through local manifold expansion.
- We propose a principled spectral entropy stopping criterion that detects early contamination and prevents pseudo-labeling drift under covariate overlap.
- We provide theoretical insights into why spectral entropy signals help prevent annotation contamination.
- We build new covariate shift benchmarks and demonstrate state-of-the-art performance, approaching fully supervised methods, while showing strong transferability across diverse shifts.

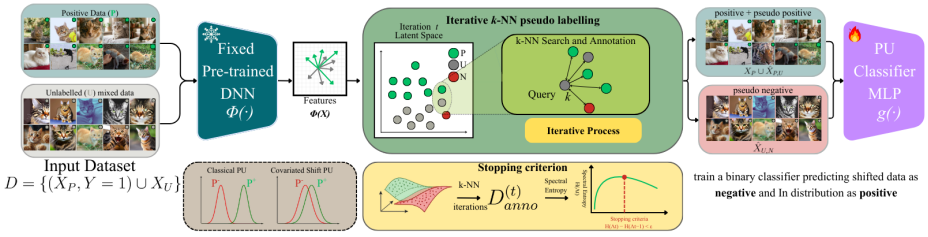

Fig. 1: Overview of S-PUNA. Given a PU dataset D, we use a pre-trained DNN $\Phi(\cdot)$ that maps samples to the latent feature space $\Phi(X)$. **S-PUNA** then performs *iterative k-NN pseudo-labeling*. k-NN expands both pseudo-positive $\tilde{X}_{P,U}$ and pseudo-negative $\tilde{X}_{U,N}$ samples symmetrically. The process is stopped using the spectral-entropy criterion. The PU classifier $g(\cdot)$ is trained to separate positives and negatives.

2 Related Works

Domain generalization. Domain adaptation (DA) [1] and domain generalization (DG) [42] also study changes in distribution, but their objective is fundamentally different from ours. In DA/DG, covariate shift is usually treated as a nuisance: the goal is to learn representations or classifiers that generalise and remain accurate and invariant across domains. In contrast, our goal is not to adapt to the shifted distribution or to improve task accuracy in the event of a shift; rather, it is the shift itself that is the target to be detected. This distinction is important because methods designed for DA/DG may deliberately remove

or obscure domain-specific information, whereas our method must preserve and exploit this information to detect the presence of a shift.

Covariate Shift Detection. While the problem is often included within broader discussions of general OOD detection [38], isolating and detecting these non-semantic shifts remains an under-explored challenge. Fully supervised strategies like DiscoPatch [6] require the unrealistic assumption that the specific covariate shifts are represented in the training set. Meanwhile, standard OOD algorithms like [2] and spatial anomaly detection methods like PatchCore [32] and SimpleNet [25] excel in semantic or localized tasks but prove ineffective for image-level covariate shift detection (see Appendix F for detailed results).

Positive-Unlabeled (PU) learning. PU learning enables binary classification using only labeled positives and an unlabeled mixture [3]. Modern approaches rely on disambiguation-free empirical risk estimators. For instance, nnPU [21] prevents overfitting by enforcing non-negative risk constraints, DCPU [24] handles biased or disjoint positive sets, and DistPU [41] mitigates negative-prediction bias by aligning predictions with the class prior through entropy minimization. PU formulations have also been applied to open-world OOD detection by treating nominal data as positives and deployment data as an unlabeled mixture [5, 30]. Most existing methods assume a semantic shift between the P and N, which simplifies the PU framework. While some works [7, 22, 33] address PU under covariate shift, they focus on classifier generalization between train and test sets rather than using the PU framework for shift detection.

3 Preliminaries and mathematical Insights

3.1 Distribution Shift Detection

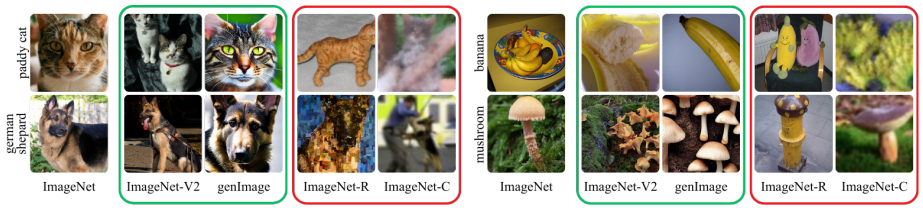

Fig. 2: Covariate shifts in ImageNet benchmarks. Comparison of natural images (ImageNet, ImageNet-V2) against synthetic corruptions (ImageNet-C), stylized rendering (ImageNet-R), and generative data (GenImage) for the classes Paddy cat (Near shifts are represented in green while Far shifts are in red).

Notations and preliminaries. Let $\mathcal{X}$ be the set of all possible inputs (usually points in $\mathbb{R}^d$, i.e. d-dimensional vectors). Let $\mathcal{Y} = \{1, 2, \ldots, K\}$ be the set of possible class labels (there are K different classes). During training we receive a dataset $\mathcal{D}_{\text{train}} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ where each pair (x_i, y_i) is supposed to come from the same joint distribution $P_{\text{train}}(X, Y)$. We train a

classifier f_θ by minimizing the average training loss:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell\big(f_\theta(x_i),\ y_i\big) \tag{1}$$

Here $\ell(\cdot, \cdot)$ denotes a loss function. Classical learning theory assumes that future test samples (X, Y) follow the same distribution: $P_{\text{test}}(X, Y) = P_{\text{train}}(X, Y)$. In practice this assumption rarely holds, leading to **distribution shift**. Detecting and handling such shifts is essential for reliable machine learning systems. Two main types are commonly distinguished: **Semantic shift** and **Covariate shift**.

Semantic Shift occurs when the label distribution differs between training and test data, i.e., $P_{\text{test}}(Y) \neq P_{\text{train}}(Y)$. This setting is closely related to *Out-of-Distribution (OOD) detection*, where test samples are drawn from a distribution P_{test} defined on the same input space $\mathcal{X}$ but involving unseen semantic classes. The OOD literature typically distinguishes two regimes [37]: **Far OOD** where the new classes are semantically very different from those seen during training, and **Near OOD** where they remain semantically similar. Near-OOD detection is generally more challenging due to the stronger semantic overlap with the training classes.

Covariate Shift refers to a change in the input distribution while the conditional labeling function remains unchanged: $P_{\text{test}}(X) \neq P_{\text{train}}(X), \quad P_{\text{test}}(Y \mid X) = P_{\text{train}}(Y \mid X)$. In this case, the optimal classifier is theoretically unchanged, but empirical risk minimization may become biased since the training data no longer reflect the marginal distribution encountered at test time. As with semantic shift, one can distinguish between **Near Shift**, where $P_{\text{test}}(X)$ differs only slightly from $P_{\text{train}}(X)$, and **Far Shift** corresponding to larger changes in input statistics. This distinction primarily reflects the difficulty of the shift rather than a strict quantitative measure in $\mathcal{X}$.

3.2 Background and Mathematical Insights

Classical Binary Classification. Let $(X, Y) \sim P$ be a random pair taking values in $\mathcal{X} \times \{-1, +1\}$, where P denotes the joint data-generating distribution. We denote by $\pi_p = \mathbb{P}(Y = +1)$, and $\pi_n = \mathbb{P}(Y = -1) = 1 - \pi_p$, the class priors, and by $P^+ = P_{X|Y=+1}$ and $P^- = P_{X|Y=-1}$, the class-conditional distributions.

Given a hypothesis class $\mathcal{H}$ of measurable functions $f : \mathcal{X} \to \mathbb{R}$ and a loss function $\ell : \mathbb{R} \times \{-1, +1\} \to [0, 1]$, the *population (true) risk* of a classifier f is defined as

$$R(f) = \mathbb{E}_{(X,Y)\sim P}[\ell(f(X), Y)]. \tag{2}$$

Using the law of total expectation, this can be decomposed as

$$R(f) = \pi_p \, \mathbb{E}_{X\sim P^+}[\ell(f(X), +1)] + \pi_n \, \mathbb{E}_{X\sim P^-}[\ell(f(X), -1)]. \tag{3}$$

In practice, P is unknown and we observe an *i.i.d.* training sample such that $D = \{(x_i, y_i)\}_{i=1}^n \sim P^n$.

The *empirical risk* is defined as

$$\hat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i). \tag{4}$$

Empirical Risk Minimization (ERM) selects $\hat{f} \in \arg\min_{f\in\mathcal{H}} \hat{R}_n(f)$.

Positive–Unlabeled (PU) Classification In PU learning, we do not observe labeled negative examples. Instead, we observe: a labeled positive set $\mathcal{X}_P = \{x_i^p\}_{i=1}^{n_p} \sim (P^+)^{n_p}$, and an unlabeled set $\mathcal{X}_U = \{x_j^u\}_{j=1}^{n_u} \sim (P_X)^{n_u}$, where the marginal distribution satisfies

$$P_X = \pi_p P^+ + \pi_n P^-. \tag{5}$$

Similarly to [14] a key observation is that the negative component of the risk can be recovered from the mixture structure. Indeed, by linearity of expectation and (5),

$$\mathbb{E}_{X\sim P_X}[\ell(f(X),-1)] = \pi_p \mathbb{E}_{X\sim P^+}[\ell(f(X),-1)] + \pi_n \mathbb{E}_{X\sim P^-}[\ell(f(X),-1)]. \tag{6}$$

Rearranging yields:

$$\pi_n \mathbb{E}_{X\sim P^-}[\ell(f(X),-1)] = \mathbb{E}_{X\sim P_X}[\ell(f(X),-1)] - \pi_p \mathbb{E}_{X\sim P^+}[\ell(f(X),-1)]. \tag{7}$$

Substituting (7) into (3) gives the *unbiased PU risk*:

$$R_{pu}(f) = \pi_p \mathbb{E}_{X\sim P^+}\big[\ell(f(X),+1) - \ell(f(X),-1)\big] + \mathbb{E}_{X\sim P_X}[\ell(f(X),-1)]. \tag{8}$$

Importantly, $R_{pu}(f) = R(f)$ for all f, hence minimizing R_{pu} is equivalent to minimizing the true risk.

An empirical estimator of (8) is

$$\hat{R}_{pu}(f) = \frac{\pi_p}{n_p}\sum_{i=1}^{n_p} \tilde{\ell}(f(x_i^p)) + \frac{1}{n_u}\sum_{j=1}^{n_u} \ell(f(x_j^u),-1), \text{ where} \tag{9}$$

$$\tilde{\ell}(f(x)) = \ell(f(x),+1) - \ell(f(x),-1). \tag{10}$$

Rademacher Generalization Bound Let $\mathcal{H}$ be a class of functions taking values in $[0,1]$. We can define its empirical Rademacher complexity:

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(x_i)\right], \text{ where } \sigma_i \text{ are i.i.d. Rademacher variables.} \tag{11}$$

A standard uniform convergence result (see, e.g., Theorem 3.1 of [28]) states that for any $\delta > 0$, with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{H}} |R(f) - \hat{R}_n(f)| \le 2\mathcal{R}_n(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2n}}. \tag{12}$$

This inequality provides a distribution-dependent control of the generalization gap.

Applying (12) separately to the positive and unlabeled samples yields, with probability at least $1-\delta$,

$$R_{pu}(f) \leq \hat{R}_{pu}(f) + 2\pi_p \mathcal{R}_{n_p}(\mathcal{H}) + 2\mathcal{R}_{n_u}(\mathcal{H}) + \pi_p \sqrt{\frac{\log(1/\delta)}{2n_p}} + \sqrt{\frac{\log(1/\delta)}{2n_u}}. \tag{13}$$

Thus, the excess risk is controlled by the complexity of the hypothesis class and the sample sizes n_p and n_u.

Bayes Optimal Risk and Total Variation Assume now balanced classes:

$$\pi_p = \pi_n = \frac{1}{2}. \tag{14}$$

Define the regression function $\eta(x) = \mathbb{P}(Y = +1 \mid X = x)$.

We can define the Bayes classifier, as in [4], by $f^*(x) = \mathbf{1}\{\eta(x) \geq 1/2\}$, which minimizes $R(f)$ over all measurable classifiers.

Its risk is

$$R(f^*) = \mathbb{E}[\min(\eta(X), 1-\eta(X))]. \tag{15}$$

Using the identity

$$\min(a,b) = \frac{1}{2}(a+b-|a-b|), \tag{16}$$

and the fact that

$$\|P^+ - P^-\|_{TV} = \frac{1}{2}\int |p^+(x) - p^-(x)|dx, \tag{17}$$

one obtains the classical result from statistical decision theory:

$$R(f^*) = \frac{1}{2}\left(1 - \|P^+ - P^-\|_{TV}\right). \tag{18}$$

Equation (18) shows that the intrinsic difficulty of the classification problem is entirely governed by the total variation distance between the class-conditional distributions. If the distributions overlap significantly, $\|P^+ - P^-\|_{TV}$ is small and the Bayes error is large. If they are well separated, $\|P^+ - P^-\|_{TV}$ approaches 1 and the Bayes error approaches 0.

3.3 Implications for Covariate Shift Detection

Under covariate shift, the marginal distribution of X changes while $P(Y \mid X)$ remains fixed. When P^+ and P^- differ substantially, the total variation term in (18) captures this shift directly.

Combining (13) with (18), we obtain the high-probability bound

$$R_{pu}(\hat{f}) \leq \frac{1}{2}\left(1 - \|P^+ - P^-\|_{TV}\right) + 2\pi_p \mathcal{R}_{n_p}(\mathcal{H}) + 2\mathcal{R}_{n_u}(\mathcal{H}) + \pi_p \sqrt{\frac{\log(1/\delta)}{2n_p}} + \sqrt{\frac{\log(1/\delta)}{2n_u}}. \tag{19}$$

This inequality highlights two sources of difficulty: **Intrinsic difficulty**, governed by $\|P^+ - P^-\|_{TV}$, which reflects distributional overlap and the **estimation error**, governed by Rademacher complexity and sample size.

In extreme classical shift scenarios used in PU, where P^+ and P^- have nearly disjoint supports, $\|P^+ - P^-\|_{TV} \approx 1$, and the intrinsic Bayes error is small. However, in our case $\|P^+ - P^-\|_{TV} \approx 0$, hence the covariate shift might complicate all the training.

4 Spectral-PU Neighborhood Annotation (S-PUNA)

In the covariate shift regime, the class-conditional distributions P^+ and P^- may exhibit substantial overlap in the ambient space, leading to a small total variation distance $\|P^+ - P^-\|_{TV}$. As shown in Section 3.2, the Bayes risk scales as $R(f^*) = \frac{1}{2}(1 - \|P^+ - P^-\|_{TV})$, implying that when $\|P^+ - P^-\|_{TV}$ is small, the intrinsic classification problem is difficult.

To mitigate this effect, we propose **Spectral–PU Neighborhood Annotation (S-PUNA)**, a progressive, geometry-aware pseudo-labeling framework. Rather than estimating the negative distribution globally, **S-PUNA** exploits the *local manifold structure* of the unlabeled data to iteratively uncover regions that are statistically inconsistent with the positive anchor distribution. This results in a gradual expansion of a reliable negative support estimate. The Appendix A gives a deeper theoretical overview of the soundness of our proposed approach.

4.1 Iterative k-NN-Based Pseudo-Labeling

Our method is a two-stage process as illustrated in Figure 1: an initialization phase of the k-NN, followed by symmetric manifold expansion.

Let us consider that we have two sets of pseudo-annotations, one for positive and one for unlabeled negative data, which we denote as: $\hat{\mathcal{X}}_{U,P}^{(t_0)} = \emptyset, \quad \hat{\mathcal{X}}_{U,N}^{(t_0)} = \emptyset$ for the first iteration t_0. Then for the next iterations $t = t_0 + 1, \ldots, T$, we maintain the current annotated sets:

$$\mathcal{X}_P^{(t)} = \mathcal{X}_P \cup \hat{\mathcal{X}}_{U,P}^{(t-1)}, \qquad \mathcal{X}_N^{(t)} = \hat{\mathcal{X}}_{U,N}^{(t-1)}. \tag{20}$$

Baseline and Sets Initialization. All samples features $\phi(x)$ are extracted from the sixth block of a pre-trained Vision Transformer (ViT) [13]. This specific intermediate representation is chosen for its efficacy in capturing the geometric nuances of covariate shifts (see Appendix D for a comparative analysis of feature

layer sensitivity). Our technique starts by constructing a k-nearest neighbor (k-NN) index on Positive Domain (PD) training features, $\mathcal{X}_{train}$. For each sample x in the unlabeled pool $\mathcal{X}_U$, we compute an initial distance score $s_{orig}(x)$ as the minimum distance from the sample to its k-nearest neighbors in $\mathcal{X}_{train}$. We then initialize the positive (PD) and negative (covariate) sets for the first iteration t_0 by ranking the pool according to these scores:

$$\hat{\mathcal{X}}_{U,c}^{(t_0)} = \begin{cases} \{x \in \mathcal{X}_U \mid x \text{ is among top } \alpha \text{ lowest } s_{orig}(\cdot)\}, & \text{if } c = P \\ \{x \in \mathcal{X}_U \mid x \text{ is among top } \alpha \text{ highest } s_{orig}(\cdot)\}, & \text{if } c = N \end{cases} \tag{21}$$

where α represents a small number of features used to seed the initial pseudo-labeled sets. For subsequent iterations $t = t_0 + 1, \ldots, T$, the current annotated sets are maintained as: $\mathcal{X}_P^{(t)} = \mathcal{X}_P \cup \hat{\mathcal{X}}_{U,P}^{(t-1)}$, $\mathcal{X}_N^{(t)} = \hat{\mathcal{X}}_{U,N}^{(t-1)}$ and $\mathcal{X}_U^{(t)} = \mathcal{X}_U^{(t-1)} / (\hat{\mathcal{X}}_{U,N}^{(t)} \cap \hat{\mathcal{X}}_{U,P}^{(t)})$.

Neighborhood Search and Scoring. At each iteration t, we compute two separate distance metrics for each candidate sample $x \in \mathcal{X}_U^{(t)}$:

1. $d_P(x)$: the mean distance from the sample features $\phi(x)$ to its k-nearest neighbors in the positive bank $\mathcal{X}_P^{(t)}$.
2. $d_N(x)$: the mean distance from the sample features $\phi(x)$ to its k-nearest neighbors in the negative bank $\mathcal{X}_N^{(t)}$.

Using these distances we calculate a score for ranking:

$$s_t(x) = d_{ood}(x) - d_{id}(x). \tag{22}$$

Set Expansion. The set is progressively expanded by ranking the candidates based on $s_t(x)$. We select the top β samples for each class to be added to the respective sets:

$$\hat{\mathcal{X}}_{U,c}^{(t)} = \hat{\mathcal{X}}_{U,c}^{(t-1)} \cup \begin{cases} \{x \mid x \text{ has top } \beta \text{ highest } s_t(x)\}, & \text{if } c = P \\ \{x \mid x \text{ has top } \beta \text{ lowest } s_t(x)\}. & \text{if } c = N \end{cases} \tag{23}$$

β is a fixed parameter we set during the whole algorithm that controls the number of samples added to the sets at each iteration.

This iterative expansion differs from classical k-NN-based PU approaches, which typically only propagate positive labels. Here, both classes are progressively constructed, leading to a symmetric refinement of the decision boundary.

Geometric Interpretation. Under the cluster assumption, points belonging to the same class lie on a low-dimensional manifold. As iterations proceed, the negative set expands along directions of maximal discrepancy from the positive support. This progressively increases the effective separation between the empirical supports of P^+ and P^-.

4.2 Spectral Entropy Stopping Criterion

A key challenge in iterative pseudo-labeling is preventing semantic drift. To detect when the discovered negative set has saturated its intrinsic structure, we monitor its spectral complexity.

Let Σ_t denote the empirical covariance matrix of $\phi(x)$ for $x \in \hat{\mathcal{X}}_{U,N}^{(t)}$:

$$\Sigma_t = \frac{1}{|\hat{\mathcal{X}}_{U,N}^{(t)}|} \sum_{x \in \hat{\mathcal{X}}_{U,N}^{(t)}} (\phi(x) - \mu_t)(\phi(x) - \mu_t)^\top, \tag{24}$$

where μ_t is the mean of the features $\phi(x)$ for $x \in \hat{\mathcal{X}}_{U,N}^{(t)}$. Let $\{\lambda_1^{(t)}, \ldots, \lambda_d^{(t)}\}$ be its eigenvalues normalized such that $\sum_i \lambda_i^{(t)} = 1$.

We define the *spectral entropy*:

$$H(\Lambda_t) = -\sum_{i=1}^{d} \lambda_i^{(t)} \log \lambda_i^{(t)}. \tag{25}$$

Interpretation. Spectral entropy quantifies the dispersion of variance across principal directions. Low entropy indicates concentration on a narrow manifold; high entropy reflects a richer, more isotropic structure. As the negative manifold is progressively uncovered, $H(\Lambda_t)$ increases.

We stop when $H(\Lambda_t) < H(\Lambda_{t-1})$, indicating that the geometric expansion has stabilized. This prevents the algorithm from invading regions where P^+ and P^- overlap, which would artificially decrease $\|P^+ - P^-\|_{TV}$.

Theoretical Justification. This stopping criterion prevents semantic drift. As formalized in Lemma 1, this inevitably disperses the mixture's spectrum and strictly decreases spectral entropy.

Lemma 1 (Spectral Entropy Decrease). *Let $\Sigma_N, \Sigma_P \succ 0$ be arbitrary within-class covariance matrices, and let μ_P and μ_N be arbitrary means real vectors. Let us assume that $\Delta\mu = \mu_P - \mu_N \neq 0$. For a contaminated mixture $X \sim (1-\varepsilon)\mathcal{N}(\mu_N, \Sigma_N) + \varepsilon\mathcal{N}(\mu_P, \Sigma_P)$, its spectral entropy satisfies $S(\varepsilon) < S(0)$ for all sufficiently small contamination levels $\varepsilon > 0$.*

The full proof is provided in Appendix A.

4.3 Knowledge Distillation into Deep Networks

After convergence, we obtain the pseudo-labeled dataset $\mathcal{D}_{anno} = \mathcal{X}_P^{(t)} \cup \hat{\mathcal{X}}_{U,N}^{(t)}$.

We train a parametric classifier g_ω on top of embeddings $\phi(x)$, e.g., a multilayer perceptron. The training objective is:

$$\mathcal{L}(\omega) = -\frac{1}{|\mathcal{D}_{anno}|} \sum_{(x,\hat{y}) \in \mathcal{D}_{anno}} \mathrm{CE}\big(g_\omega(\phi(x)), \hat{y}\big). \tag{26}$$

Why Distillation Helps. The k-NN procedure is a non-parametric estimator with low bias but high variance. Distilling its pseudo-labels into a DNN induces a smoother decision boundary, reducing variance while preserving local geometry. Importantly, the learned DNN is optimized to keep the local geometry that comes from the k-NN, allowing for a good separation between positive and negative.

5 Experimental results

In this section, we evaluate our proposed method **S-PUNA** and test its robustness under severe covariate shift across diverse datasets. We benchmark **S-PUNA** against classical positive-unlabeled (PU) risk minimization estimators as well as neighborhood-based baseline methods. Complete implementation details and hyperparameters are provided in the Appendix B.3. In Section 5.3, we analyze 21 methods using ResNet and ViT architectures to evaluate how supervised and unsupervised techniques perform in detecting covariate shift. We provide some examples of downstream applications of **S-PUNA** in Appendix I.

5.1 Datasets and Metrics

We evaluate our proposed method across three primary image classification benchmarks characterized by significant covariate shifts: **ImageNet-1K** [10], **TinyImageNet** [10], and **EuroSAT** [16].

ImageNet-1K Benchmarks. Using ImageNet-1K as the Positive Domain (PD) baseline, we test our approach on four shifted datasets. **ImageNet-V2** [31] (near shift) replicates the original data collection process while introducing subtle statistical biases. **ImageNet-R** [17] (far shift) provides adversarial and rendered versions (e.g., paintings, sketches) of ImageNet classes, preserving semantic content while altering the visual domain. **ImageNet-C** (far shift) [18] applies 15 types of algorithmic perturbations, including noise, blur, and weather effects, across five severity levels. Finally, **GenImage** (near shift) [44] consists of ImageNet classes regenerated via various generative models, introducing synthetic distributional shifts.

TinyImageNet Benchmarks. For TinyImageNet (PD), we consider two shifted variants: **TinyImageNet-C** [18] (far shift), which applies corruptions similar to ImageNet-C, and **TinyImageNet-V2** (near shift), a filtered version of ImageNet-V2 containing only the 200 overlapping classes.

EuroSAT Benchmarks. To demonstrate the versatility of our method beyond object-centric domains, we evaluate it on the EuroSAT satellite dataset (PD). We consider two shifted variants: **EuroSAT-C**(far shift) [29], which simulates synthetic atmospheric and sensor-based degradations, and a new dataset introduced in this study, **EuroSAT-D**(near shift). The latter consists of regenerated versions of EuroSAT images to address synthetic generative shifts. Further details on dataset construction are provided in Appendix B.4 .

Metrics and Evaluation. We report standard covariate shift detection metrics: Area Under the Receiver Operating Characteristic curve (**AUROC**), Area

Under the Precision-Recall curve (**AUPR**), and the False Positive Rate at a 95% True Positive Rate (**FPR95**).

In our analysis, we distinguish between **near-shift** datasets, which exhibit moderate covariate shifts that are distributionally close to the PD data (and thus more challenging to detect), and **far-shift** datasets, which represent severe shifts that are more easily identifiable. We provide a comprehensive discussion on this taxonomy in Appendix H.

5.2 Baselines and Implementation Details

For evaluation, images are projected into a dense feature space extracted from the sixth block of a ViT pre-trained on in-domain data. We select this specific intermediate layer as it captures structural nuances of covariate shifts while mitigating the task-specific overfitting prevalent in deeper layers (see Appendix D).

We benchmark **S-PUNA** against several state-of-the-art PU learning methods, including **Dist-PU** [41], **DC-PU** [24], and **saPU** [9], all of which are optimized using the same ViT feature embeddings. Additionally, we compare our approach against pseudo-labeling strategies such as **LaGAM** [26] and a **naive k-NN baseline** [40] that propagates labels from the positive set.

For S-PUNA, the iterative expansion process terminates when the absolute difference in spectral entropy between consecutive steps, $H(\Lambda_t) - H(\Lambda_{t-1})$, falls below the threshold of 0.0. An ablation study regarding all of the hyperparameters of **S-PUNA** is provided in Appendix C. Upon convergence, the learned structural information is distilled into an MLP via standard cross-entropy training. Results using a ResNet instead of the ViT are detailed in Appendix G.

5.3 From supervised to unsupervised Covariate shift detection

To assess the effectiveness of various detection paradigms, we conduct an extensive benchmark in Appendix F, evaluating 14 OOD detection methods, 5 anomaly detection methods, and 3 diffusion-based techniques across both ResNet and ViT architectures. We further compare these against two supervised approaches: *DiscoPatch* [6] and *DRCT* [8]. Our results indicate that supervised techniques achieve superior performance on **near-shift** scenarios, with an average AUROC of 89.8%. However, on **far-shift** datasets, these methods are outperformed by the majority of unsupervised techniques. Among the unsupervised approaches, logit-based methods such as *ReAct* [34] and *ASH* [12] yield the weakest results, suggesting that high-level class probabilities often discard the subtle structural cues necessary for robust

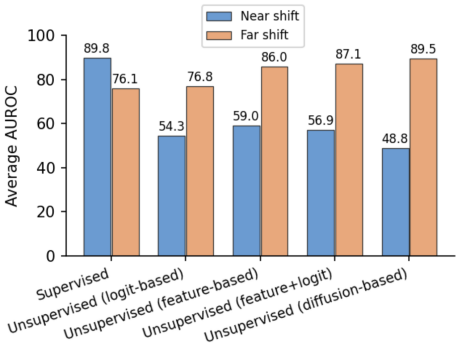

Fig. 3: Overview of the performance of ood detection methods on covariate shifted datasets of imagenet1k

detection. While feature-based methods like *MDS* [23], hybrid approaches like *ViM* [36] and diffusion based methods like DiffPath [19] show marginal improvements, these findings highlight that current OOD detection methods struggle on covariate shift when restricted to purely unsupervised internal signals.

5.4 Results

ImageNet. Table 1 summarizes the covariate shift detection performance on ImageNet-1K for both near and far shift scenarios.

Classical PU-based baselines exhibit a significant performance gap between the two regimes. For instance, **Dist-PU** improves from 84.48% AUROC in near-shift cases to 96.70% in far-shift cases, confirming that the latter is considerably easier to identify. However, its near-shift performance remains insufficient, with a high FPR95 of 35.05%. Similar trends are observed for **saPU** and **LaGAM**, while **DC-PU** struggles across both regimes, particularly regarding FPR95 (85.80% near, 89.38% far).

In contrast, **S-PUNA** consistently outperforms all baselines. Under near-shift conditions, our method achieves 97.89% AUROC—surpassing the strongest baseline (**Dist-PU**) by over 13%—while simultaneously reducing the FPR95 from 35.05% to 9.69% (a −25.3% improvement). **S-PUNA** also achieves the highest AUROC under far-shift (98.51%). On average, our approach yields 98.20% AUROC (a +7.61% gain over **Dist-PU**) and reduces the average FPR95 to 8.52%, outperforming the best baseline by nearly 18%. Unlike prior methods, our approach maintains high stability across both shift types, demonstrating superior robustness.

Table 1: Comparison of methods on **Imagenet** with **Near Covariate Shift** and **Far Covariate Shift** (**w/ C**: with classifier; **w/o C**: without classifier; **C**: classifier). We report AUROC, AUPR, and FPR95 (%). Results are shown as *mean* ± *std*.

Method	Near Shift			Far Shift			Average		
	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓
k-NN [40]	54.40±0.00	90.92±0.00	62.63±0.00	74.79±0.00	60.78±0.00	80.84±0.00	64.59±0.00	75.85±0.00	71.73±0.00
Dist-PU [41]	84.48±0.40	87.33±0.30	35.05±0.93	96.70±0.11	96.51±0.14	18.06±0.58	90.59±0.25	91.92±0.22	26.56±0.75
saPU [9]	82.93±0.27	85.83±0.26	69.58±1.01	95.09±0.46	87.53±3.17	26.78±2.33	89.01±0.37	86.68±1.72	48.18±1.67
Dc-PU [24]	60.61±2.44	65.26±1.70	85.80±1.03	70.51±3.96	41.98±4.34	89.38±10.83	65.56±3.20	53.62±3.02	87.59±5.93
LaGAM [26]	83.01±0.54	86.40±0.56	65.30±1.96	96.50±0.13	94.36±1.66	17.46±0.71	89.75±0.33	90.38±1.11	41.38±1.33
S-PUNA w/o C (ours)	95.81±0.00	97.66±0.00	19.22±0.00	95.95±0.00	97.93±0.00	19.92±0.00	95.88±0.00	97.79±0.00	19.57±0.00
S-PUNA w/ C (ours)	**97.89±0.13**	**98.81±0.08**	**9.69±0.48**	**98.51±0.08**	**99.20±0.05**	**7.36±0.49**	**98.20±0.10**	**99.00±0.06**	**8.52±0.48**

TinyImageNet. On the TinyImageNet benchmark (Table 2), most baselines perform exceptionally well under near-shift conditions; for instance, **Dist-PU** achieves 99.96% AUROC and 0.15% FPR95. **S-PUNA** surpasses these results, reaching perfect near-shift performance (100.00% AUROC, 0.00% FPR95). Our method also provides the strongest far-shift results (99.64% AUROC, 1.20% FPR95), slightly improving upon the best-performing baseline, **LaGAM** (99.59% AUROC, 1.39% FPR95). On average, **S-PUNA** consistently improves all metrics

Table 2: Comparison of methods on **Tiny-ImageNet** with **Near Covariate Shift** and **Far Covariate Shift** (**w/ C**: with classifier; **w/o C**: without classifier; **C**: classifier). We report AUROC, AUPR, and FPR95 (%). Results are shown as *mean* ± *std*.

Method	Near Shift			Far Shift			Average		
	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓
k-NN [40]	82.00±0.00	50.65±0.00	78.39±0.00	95.82±0.00	28.19±0.00	96.96±0.00	88.91±0.00	39.42±0.00	87.68±0.00
Dist-PU [41]	99.96±0.00	99.98±0.00	0.15±0.04	99.46±0.11	96.14±0.62	2.81±0.60	99.71±0.06	98.06±0.31	1.48±0.32
saPU [9]	91.73±0.01	96.37±0.01	19.14±0.01	92.07±0.17	71.47±1.97	33.44±0.40	91.90±0.09	83.92±0.99	26.29±0.21
Dc-PU [24]	95.93±2.90	97.82±1.52	16.43±13.11	82.95±5.58	35.21±14.85	61.76±21.21	89.44±4.24	66.52±8.19	39.09±17.16
LaGAM [26]	99.93±0.01	99.96±0.01	0.22±0.01	99.59±0.21	95.35±3.03	1.39±0.78	99.76±0.11	97.66±1.52	0.81±0.39
S-PUNA w/o C (ours)	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	99.54±0.00	**99.62±0.00**	**1.20±0.00**	**99.77±0.00**	**99.81±0.00**	**0.60±0.00**
S-PUNA w/ C (ours)	99.90±0.10	99.92±0.08	0.02±0.04	**99.64±0.11**	99.46±0.15	1.51±0.50	**99.77±0.11**	99.69±0.12	0.77±0.27

across the board compared to the state-of-the-art.

EuroSAT. Results on the EuroSAT dataset (Table 3) reveal a critical vulnerability in traditional PU methods. While several baselines achieve high AUROC under near-shift conditions (**Dist-PU**: 99.78%), their performance degrades substantially under far-shift scenarios—**Dist-PU**, for example, drops to 89.46% AUROC with a significant 49.97% FPR95. In contrast, **S-PUNA** achieves perfect separation in the far-shift regime (100.00% AUROC, 0.00% FPR95), marking a +10.54% AUROC gain and a −49.97% reduction in FPR95 over the best baseline. Averaged across both regimes, **S-PUNA** reaches 99.90% AUROC and 0.33% FPR95, substantially outperforming **LaGAM**'s average metrics (92.79% AUROC, 28.92% FPR95). These results confirm the superior robustness of our geometry-aware approach across varying shift severities. Please refer to Appendix E for transfer experiments and robustness across shifts.

Table 3: Comparison of methods on **EuroSAT** with **Near** and **Far Covariate Shift**. (**w/ C**: with classifier; **w/o C**: without classifier; **C**: classifier). We report AUROC, AUPR, and FPR95 (%). Results are shown as *mean* ± *std*.

Method	Near Shift			Far Shift			Average		
	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓	AUROC↑	AUPR↑	FPR95↓
k-NN [40]	58.86±0.00	84.80±0.00	13.91±0.00	81.03±0.00	83.11±0.00	58.47±0.00	69.95±0.00	83.96±0.00	36.19±0.00
Dist-PU [41]	99.78±0.13	**99.95±0.03**	0.72±0.46	89.46±0.50	96.48±0.26	49.97±2.81	94.62±0.31	**98.22±0.14**	25.35±1.64
saPU [9]	96.71±0.93	99.30±0.21	7.55±3.58	86.75±0.62	95.81±0.15	54.38±1.06	91.73±0.78	97.56±0.18	30.97±2.32
Dc-PU [24]	68.96±3.53	87.87±0.59	90.89±6.73	60.22±4.34	85.05±1.43	94.48±3.06	64.59±3.93	86.46±1.01	92.68±4.90
LaGAM [26]	99.33±0.56	99.86±0.14	3.01±3.39	86.24±0.90	96.28±0.27	54.83±3.24	92.79±0.73	98.07±0.21	28.92±3.32
S-PUNA w/o C (ours)	84.40±0.00	38.02±0.00	55.76±0.00	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	92.20±0.00	69.01±0.00	27.88±0.00
S-PUNA w/ C (ours)	**99.79±0.05**	96.08±0.70	**0.65±0.36**	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**99.90±0.03**	98.04±0.35	**0.33±0.18**

5.5 Ablation and Sensitivity Analysis

Classifier head contribution. Across all datasets, we observe that the classifier head often improves over the S-PUNA w/o C variant; when it does not, its performance remains comparable within the metric ± the standard deviation.

Effect of the proportion of shifted data. Previous experiments considered the standard 1:1 composition of the unlabeled pool, corresponding to **50%** shifted samples in Table 4. Using the same features extracted from block 6 of the ViT, we conducted additional experiments on ImageNet-1K and its associated near- and

Table 4: AUROC across different shifted-sample ratios in INet1K.

Method	Average AUROC				
	Block 11	Block 6			
		$p = 5\%$	$p = 25\%$	$p = 50\%$	$p = 75\%$
DCPU	62.23	73.08	68.00	65.56	72.19
DistPU	82.08	71.04	86.40	90.59	92.17
LaGAM	79.35	85.08	87.45	89.75	90.92
SAPU	79.38	60.24	75.53	89.01	92.03
S-PUNA w/c	**97.03**	**91.61**	**95.75**	**98.20**	**95.53**

far-shifted datasets with shifted-sample proportions of **5%**, **25%**, and **75%**. Table 4 reports the average AUROC for covariate-shift detection across all datasets. S-PUNA consistently outperforms the PU-learning baselines for every considered proportion of shifted data.

Effect of ViT layer selection. We also compared features extracted from blocks 6 and 11 of the ViT for shift detection on ImageNet-1K, while keeping the S-PUNA pipeline and all hyperparameters unchanged. The corresponding results are reported in the first column of Table 4. S-PUNA continues to outperform the PU-learning baselines when block-11 features are used, indicating that its improvement is not merely an artifact of selecting block 6. Nevertheless, block 6 provides the best overall performance. Additional results are in Appendix D.

6 Conclusion

In this work, we investigated the efficacy of various paradigms for covariate shift detection. Our analysis demonstrates that while fully supervised approaches perform reasonably well, traditional unsupervised covariate shift detection methods—relying on internal signals like logits or feature statistics fail significantly in the near shift regime. Because covariate shifts preserve semantic content and introduce only subtle distributional changes, these unsupervised signals are unable to reliably distinguish shifted samples from in-distribution data.

To bridge this gap, we proposed a weakly supervised alternative based on the Positive–Unlabeled (PU) learning framework. To mitigate the instability of classical PU risk estimators under significant distributional overlap, we introduced **Spectral PU Neighborhood Annotation (S-PUNA)**. This geometry-aware method progressively discovers shifted samples through local neighborhood expansion within the feature manifold.

Extensive experiments across multiple benchmarks confirm that **S-PUNA** consistently outperforms existing PU baselines and even surpasses fully supervised models in several settings. Our findings suggest that strictly supervised signals are not a prerequisite for reliable covariate shift detection; rather, by effectively leveraging the intrinsic geometric structure of feature representations, weakly supervised PU learning can achieve highly robust and state-of-the-art performance.

Acknowledgments. This work was granted access to the HPC resources of IDRIS under the allocation AD011015965R1 and AD011016949 made by GENCI.

References

1. Ajith, A., Gopakumar, G.: Domain adaptation: A survey. In: Computer Vision and Machine Intelligence: Proceedings of CVMI 2022, pp. 591–602. Springer (2023)
2. Ammar, M.B., Belkhir, N., Popescu, S., Manzanera, A., Franchi, G.: Neco: Neural collapse based out-of-distribution detection. arXiv preprint arXiv:2310.06823 (2023)
3. Bekker, J., Davis, J.: Learning from positive and unlabeled data: a survey. Machine Learning **109**(4), 719–760 (2020)
4. Bishop, C.M., Nasrabadi, N.M.: Pattern recognition and machine learning, vol. 4. Springer (2006)
5. Blanchard, G., Lee, G., Scott, C.: Semi-supervised novelty detection. The Journal of Machine Learning Research **11**, 2973–3009 (2010)
6. Caetano, F., Viviers, C., Zavala-Mondragón, L.A., de With, P.H., van der Sommen, F.: Discopatch: Taming adversarially-driven batch statistics for improved out-of-distribution detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
7. Chan, A., Alaa, A., Qian, Z., Van Der Schaar, M.: Unlabelled data improves bayesian uncertainty calibration under covariate shift. In: International conference on machine learning. pp. 1392–1402. PMLR (2020)
8. Chen, B., Zeng, J., Yang, J., Yang, R.: Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In: Forty-first International Conference on Machine Learning (2024)
9. Dai, Y., Hou, Z., Li, C., Xu, Y., Wang, E., Li, X.: A closer look to positive-unlabeled learning from fine-grained perspectives: An empirical study. In: Advances in Neural Information Processing Systems (NeurIPS) (2025), `https://openreview.net/forum?id=gioV0q55AJ`, neurIPS 2025 (poster)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
11. Dixon, M.F., Halperin, I., Bilokon, P., et al.: Machine learning in finance, vol. 1170. Springer (2020)
12. Djurisic, A., Bozanic, N., Ashok, A., Liu, R.: Extremely simple activation shaping for out-of-distribution detection. In: The Eleventh International Conference on Learning Representations (2022)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
14. Du Plessis, M.C., Niu, G., Sugiyama, M.: Analysis of learning from positive and unlabeled data. Advances in neural information processing systems **27** (2014)
15. Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I.S., Saria, S.: The clinician and dataset shift in artificial intelligence. New England Journal of Medicine **385**(3), 283–286 (2021)
16. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **12**(7), 2217–2226 (2019)
17. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al.: The many faces of robustness: A critical

analysis of out-of-distribution generalization. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8340–8349 (2021)
18. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. arXiv preprint arXiv:1903.12261 (2019)
19. Heng, A., Thiery, A.H., Soh, H.: Out-of-distribution detection with a single unconditional diffusion model. Advances in Neural Information Processing Systems **37**, 43952–43974 (2024)
20. Katz-Samuels, J., Nakhleh, J.B., Nowak, R., Li, Y.: Training ood detectors in their natural habitats. In: International Conference on Machine Learning. pp. 10848–10865. PMLR (2022)
21. Kiryo, R., Niu, G., du Plessis, M.C., Sugiyama, M.: Positive-unlabeled learning with non-negative risk estimator. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
22. Kumagai, A., Iwata, T., Takahashi, H., Nishiyama, T., Adachi, K., Fujiwara, Y.: Positive-unlabeled auc maximization under covariate shift. In: Forty-second International Conference on Machine Learning (2025)
23. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. Advances in neural information processing systems **31** (2018)
24. Li, X., Dai, Y., Wang, B., Li, C., Qu, J., Guan, R.: Balancing positive and negative classification error rates in positive-unlabeled learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2025), `https://openreview.net/forum?id=DTq0CLhN4A`, neurIPS 2025 (poster)
25. Liu, Z., Zhou, Y., Xu, Y., Wang, Z.: Simplenet: A simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20402–20411 (2023)
26. Long, L., Wang, H., Jiang, Z., Feng, L., Yao, C., Chen, G., Zhao, J.: Positive-unlabeled learning by latent group-aware meta disambiguation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
27. Lu, J., Liu, A., Dong, F., Gu, F., Gama, J., Zhang, G.: Learning under concept drift: A review. IEEE Transactions on Knowledge and Data Engineering **31**, 2346–2363 (2019), `https://api.semanticscholar.org/CorpusID:69449458`
28. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. MIT Press, Cambridge, MA, 2nd edn. (2018), `https://mlcontroversial.com/foundations-of-machine-learning-2nd-edition/`
29. Oehri, S., Ebert, N., Abdullah, A., Stricker, D., Wasenmüller, O.: Genformer–generated images are all you need to improve robustness of transformers on small datasets. In: International Conference on Pattern Recognition. pp. 176–192. Springer (2024)
30. Pang, G., Shen, C., Van Den Hengel, A.: Deep anomaly detection with deviation networks. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 353–362 (2019)
31. Recht, B., Roelofs, R., Schmidt, L., Shankar, V.: Do imagenet classifiers generalize to imagenet? In: International conference on machine learning. pp. 5389–5400. PMLR (2019)
32. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14318–14328 (2022)

33. Sakai, T., Shimizu, N.: Covariate shift adaptation on learning from positive and unlabeled data. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 4838–4845 (2019)
34. Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. Advances in neural information processing systems **34**, 144–157 (2021)
35. Takahashi, H., Iwata, T., Kumagai, A., Yamanaka, Y.: Deep positive-unlabeled anomaly detection for contaminated unlabeled data. arXiv preprint arXiv:2405.18929 (2024)
36. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4921–4930 (2022)
37. Yang, J., Wang, P., Zou, D., Zhou, Z., Ding, K., Peng, W., Wang, H., Chen, G., Li, B., Sun, Y., et al.: Openood: Benchmarking generalized out-of-distribution detection. Advances in Neural Information Processing Systems **35**, 32598–32611 (2022)
38. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. International Journal of Computer Vision **132**(12), 5635–5662 (2024)
39. Yu, P., Xia, Z., Fei, J., Lu, Y.: A survey on deepfake video detection. Iet Biometrics **10**(6), 607–624 (2021)
40. Zhang, B., Zuo, W.: Reliable negative extracting based on knn for learning from positive and unlabeled examples. J. Comput. **4**(1), 94–101 (2009)
41. Zhao, Y., et al.: Dist-pu: Positive-unlabeled learning from a label distribution perspective. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
42. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4396–4415 (2022)
43. Zhu, M., Chen, H., Huang, M., Li, W., Hu, H., Hu, J., Wang, Y.: Gendet: Towards good generalizations for ai-generated image detection. arXiv preprint arXiv:2312.08880 (2023)
44. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: Genimage: A million-scale benchmark for detecting ai-generated image. Advances in neural information processing systems **36**, 77771–77782 (2023)