Deep Anomaly Detection under Labeling Budget Constraints

Aodong Li [*1] **Chen Qiu** [*2] **Marius Kloft** [3] **Padhraic Smyth** [1] **Stephan Mandt** [†1] **Maja Rudolph** [†2]

Abstract

Selecting informative data points for expert feedback can significantly improve the performance of anomaly detection (AD) in various contexts, such as medical diagnostics or fraud detection. In this paper, we determine a set of theoretical conditions under which anomaly scores generalize from labeled queries to unlabeled data. Motivated by these results, we propose a data labeling strategy with optimal data coverage under labeling budget constraints. In addition, we propose a new learning framework for semi-supervised AD. Extensive experiments on image, tabular, and video data sets show that our approach results in state-of-the-art semi-supervised AD performance under labeling budget constraints.

1. Introduction

Detecting anomalies in data is a fundamental task in machine learning with applications across multiple domains, from industrial fault detection to medical diagnosis. The main idea is to train a model (such as a neural network) on a data set of "normal" samples to minimize the loss of an auxiliary (e.g., self-supervised) task. Using the loss function to score test data, one hopes to obtain low scores for normal data and high scores for anomalies (Ruff et al., 2021).

In practice, the training data is often contaminated with unlabeled anomalies that differ in unknown ways from the i.i.d. samples of normal data. No access to a binary anomaly label (indicating whether a sample is normal or not) makes learning the anomaly scoring function from contaminated data challenging; the training signal has to come exclusively from the input features (typically real-valued vectors). Many approaches either assume that the unlabeled anomalies are too rarely encountered during training to affect learning (Wang et al., 2019) or try to detect and exploit the anomalies in the training data (e.g., Qiu et al. (2022a)).

While AD is typically an unsupervised training task, sometimes expert feedback is available to check if individual samples are normal or not. For example, in a medical setting, one may ask a medical doctor to confirm whether a given image reflects normal or abnormal cellular tissue. Other application areas include detecting network intrusions or machine failures. Anomaly labels are usually expensive to obtain but are very valuable to guide an anomaly detector during training. For example, in Fig. 1, we can see that our method, with only one labeled query (Fig. 1 d) is almost on par with supervised AD (Fig. 1 a). However, the supervised setting is unrealistic, since expert feedback is typically expensive. Instead, it is essential to develop effective strategies for querying informative data points.

Previous work on AD under a labeling budget primarily involves domain-specific applications and/or ad hoc architectures, making it hard to disentangle modeling choices from querying strategies (Trittenbach et al., 2021). In contrast, this paper theoretically and empirically studies generalization performance using various labeling budgets, querying strategies, and losses.

In summary, our main contributions are as follows:

1. We prove that the ranking of anomaly scores generalizes from labeled queries to unlabeled data under certain conditions that characterize how well the queries cover the data. Based on this theory, we propose a diverse querying strategy for deep AD under labeling budget constraints.
2. We propose **s**emi-supervised **o**utlier **e**xposure with a **l**imited labeling budget (SOEL), a semi-supervised learning framework compatible with a large number of deep AD losses. We show how all major hyperparameters can be eliminated, making SOEL easy to use. To this end, we provide an estimate for the anomaly ratio in the data.
3. We provide an extensive benchmark for deep AD with a limited labeling budget. Our experiments on image, tabular, and video data provide evidence that SOEL outperforms existing methods significantly. Comprehensive ablations disentangle the benefits of each component.

[*]Equal contribution [†]Joint supervision [1]Department of Computer Science, University of California, Irvine, USA [2]Bosch Center for Artificial Intelligence, Pittsburgh, USA [3]Department of Computer Science, TU Kaiserslautern, Germany. Correspondence to: Aodong Li <aodongl1@uci.edu>, Stephan Mandt <mandt@uci.edu>, Maja Rudolph <Maja.Rudolph@us.bosch.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Our paper is structured as follows. Sec. 2 introduces the problem setting we address and our main algorithm. Sec. 3 discusses related work in deep AD. Sec. 4 discusses experimental results on each of image, video, and tabular data. Finally, we conclude this work in Section 5.

2. Methods

2.1. Notation and Problem Statement

Consider a dataset $\{\boldsymbol{x}_i\}_{i=1}^N$ where the datapoints $\boldsymbol{x}_i$ are i.i.d. samples from a mixture distribution $p(\boldsymbol{x}) = (1-\alpha)p_0(\boldsymbol{x}) + \alpha p_1(\boldsymbol{x})$. The distribution $p_0(\boldsymbol{x})$ corresponds to the normal data, while $p_1(\boldsymbol{x})$ corresponds to anomalous data. We assume that $0 \leq \alpha < 0.5$, i.e., that the anomalous data is non-dominant in the mixture; in practice, $\alpha \ll 0.5$.

In the AD problem, we wish to use the data to train an anomaly detector in the form of a parametric anomaly score function $S(\boldsymbol{x};\theta)$. Once trained this score function is thresholded to determine whether a datapoint $\boldsymbol{x}_i$ is anomalous, as indicated by the binary anomaly label $y_i := y(\boldsymbol{x}_i) \in \{0 := \text{“normal”}, 1 := \text{“abnormal”}\}$.

We focus on the situation where the training data is unlabeled (only $\boldsymbol{x}_i$ is known, not y_i), but where we have access to an oracle (e.g., a human expert) that is able to provide labels y_i for a budgeted number K of the N training points.

2.2. Outline of the Technical Approach

Our work addresses the following questions: How to best select informative data points for labeling – this is called the *querying strategy*, how to best learn an anomaly detector from both the labeled and unlabeled data in a semi-supervised fashion, and how to make the approach easy to use by eliminating a crucial hyper-parameter.

Querying Strategy. A successful approach for deep AD under labeling budget constraints will require a strategy for selecting the most beneficial set of queries. We choose a theoretically-grounded approach based on generalization performance. For this, we exploit that at test-time an AD method will threshold the anomaly scores to distinguish between normal samples and anomalies. This means that the quality of a scoring function is not determined by the absolute anomaly scores but only by their relative ranking. In Sec. 2.4, we characterize a favorable property of the query set which can guarantee that the ranking of anomaly scores generalizes from the labeled data to unlabeled samples. Since this is desirable, we derive a querying strategy that under a limited labeling budget best fulfills the favorable properties put forth by our analysis.

Semi-supervised Outlier Exposure. As a second contribution, we propose a semi-supervised learning framework that best exploits both the labeled query set and the unlabeled data. It builds on supervised AD and latent outlier exposure (LOE) which we review in Sec. 2.3. We present SOEL in Sec. 2.5. The SOEL training objective is designed to receive opposing training signals from the normal samples and the anomalies. An EM-style algorithm alternates between estimating the anomaly labels of the unlabaled data and improving the anomaly scoring function using the data samples and their given or estimated labels.

Hyperparameter Elimination. Like related methods discussed in Sec. 3, SOEL has an important hyperparameter α which corresponds to the expected fraction of anomalies in the data. While previous work has to assume that α is known (Qiu et al., 2022a), our proposed method presents an opportunity to estimate it. The estimate has to account for the fact that the optimal querying strategy derived from our theory in Sec. 2.4 is not i.i.d.. In Sec. 2.6, we provide an estimate of α for any stochastic querying strategy.

2.3. Background: Deep AD

In deep AD, auxiliary losses help learn the anomaly scoring function $S(\boldsymbol{x};\theta)$. Popular losses include autoencoder-based losses (Zhou and Paffenroth, 2017), the deep SVDD loss (Ruff et al., 2018), or the neural transformation learning loss (Qiu et al., 2021). It is assumed that minimizing such a loss $L_0^\theta(\boldsymbol{x}) \equiv \mathcal{L}_0(S(\boldsymbol{x};\theta))$ over “normal” data leads to a desirable scoring function that assigns low scores to normal samples and high scores to anomalies.

Most deep AD methods optimize such an objective over an entire unlabeled data set, even if it contains unknown anomalies. It is assumed that the anomalies are rare enough that they will not dilute the training signal provided by the normal samples (*inlier priority*, (Wang et al., 2019)). Building on the ideas of Ruff et al. (2019) that synthetic anomalies can provide valuable training signal, Qiu et al. (2022a) show how to discover and exploit anomalies by treating the anomaly labels as latent variables in training.

The key idea of Ruff et al. (2019) is to construct a complementary loss $L_1^\theta(\boldsymbol{x}) \equiv \mathcal{L}_1(S(\boldsymbol{x};\theta))$ for anomalies that has an opposing effect to the normal loss $L_0^\theta(\boldsymbol{x})$. For example, the deep SVDD loss $L_0^\theta(\boldsymbol{x}) = ||f_\theta(\boldsymbol{x}) - \mathbf{c}||^2$, with feature extractor f_θ, pulls normal data points towards a fixed center c (Ruff et al., 2018). The opposing loss for anomalies, defined in Ruff et al. (2019) as $L_1^\theta(\boldsymbol{x}) = 1/L_0^\theta(\boldsymbol{x})$, pushes abnormal data away from the center.

Supervised AD. Using only the labeled data indexed by $\mathcal{Q}$ one could train $S(\boldsymbol{x};\theta)$ using a supervised loss (Hendrycks et al., 2018; Görnitz et al., 2013)

$$\mathcal{L}_{\mathcal{Q}}(\theta) = \frac{1}{|\mathcal{Q}|} \sum_{j \in \mathcal{Q}} \big(y_j L_1^\theta(\boldsymbol{x}_j) + (1-y_j) L_0^\theta(\boldsymbol{x}_j)\big). \quad (1)$$

Latent Outlier Exposure. Latent outlier exposure (LOE,

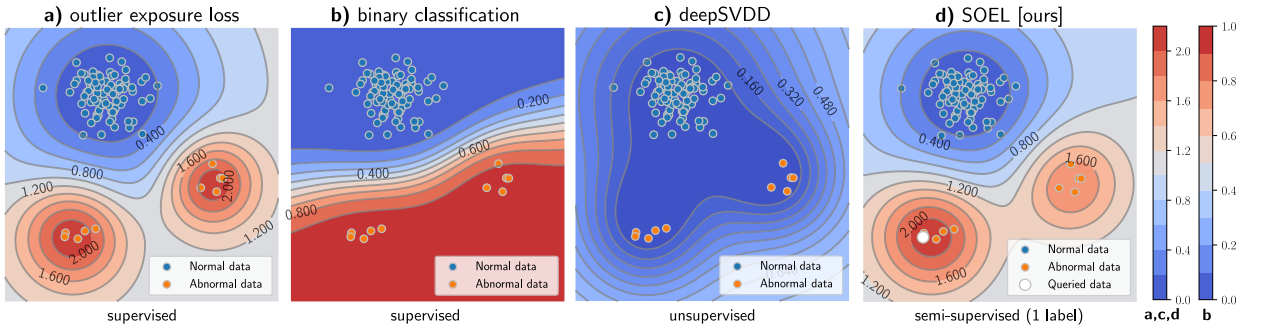

Figure 1. Anomaly score contour plots on 2D toy data demonstrate that SOEL [ours, (**d**)] with only one labeled sample can achieve detection accuracy that is competitive with a fully supervised approach (**a**). Binary classification (**b**) is problematic for AD since it cannot detect new anomalies, e.g. in the upper right corner of the plot. Subplot (**c**) demonstrates that unsupervised AD is challenging with contaminated data. Even a single labeled query, in combination with our approach, can significantly improve AD.

(Qiu et al., 2022a)) is an unsupervised AD framework that uses the same loss as Eq. (1) but treats the labels y as latent variables. An EM-style algorithm alternates between optimizing the model w.r.t. θ and inferring the labels y.

In this work, we propose semi-supervised outlier exposure with a limited labeling budget (SOEL) which builds on these ideas. We next present the querying strategy and when the querying strategy leads to correct ranking of anomaly scores (Sec. 2.4), the SOEL loss (Sec. 2.5), and how the hyperparameter α can be eliminated (Sec. 2.6)

2.4. Querying Strategies for AD

The first ingredient of SOEL is a querying strategy for selecting informative data points to be labeled, which we derive from theoretical considerations. An important property of the querying strategy is how well it covers unlabeled data. The quality of a querying strategy is determined by the smallest radius δ such that all unlabeled points are within distance δ of one queried sample of the same type. In this paper, we prove that if the queries cover both the normal data and the anomalies well (i.e., if δ is small), a learned anomaly detector that satisfies certain conditions is guaranteed to generalize correctly to the unlabeled data (The exact statement and its conditions will be provided in Thm. 1). Based on this insight, we propose to use a querying strategy that is better suited for deep AD than previous work.

Theorem 1. *Let $\mathcal{Q}_0$ be the index set of datapoints labeled normal and $\mathcal{Q}_1$ the index set of datapoints labeled abnormal. Let $\delta \in \mathbb{R}^+$ be the smallest radius, such that for each unlabeled anomaly $\boldsymbol{u}_a$ and each unlabeled normal datum $\boldsymbol{u}_n$ there exist labeled data points $\boldsymbol{x}_a, a \in \mathcal{Q}_1$ and $\boldsymbol{x}_n, n \in \mathcal{Q}_0$, such that $\boldsymbol{u}_a$ is within the δ-ball of $\boldsymbol{x}_a$ and $\boldsymbol{u}_n$ is within the δ-ball around $\boldsymbol{x}_n$. If a λ_s-Lipschitz continuous function S ranks the labeled data correctly, with a large enough margin, i.e. $S(\boldsymbol{x}_a) - S(\boldsymbol{x}_n) \geq 2\delta\lambda_s$, then S ranks the unlabeled points correctly, too, and $S(\boldsymbol{u}_a) \geq S(\boldsymbol{u}_n)$.*

In Supp. A, we prove Thm. 1 and discuss the assumptions. An implication of this theorem is that a smaller δ corresponding to a tighter cover of the data leads to better-generalized ranking performance. As detailed in Supp. A, there is a connection between correct anomaly score ranking and high AUROC performance, a common evaluation metric for AD.

Existing methods use querying strategies that do not have good coverage and are therefore not optimal under Thm. 1. For a limited querying budget, random querying puts too much weight on high-density areas of the data space, while other strategies only query locally, e.g., close to an estimated decision boundary between normal and abnormal data.

Proposed Querying Strategy. Based on Thm. 1, we propose a querying strategy that encourages tight coverage: diverse querying. In practice, we use the seeding algorithm of k-means++ which is usually used to initialize diverse clusters.[1] It iteratively samples another data point to be added to the query set $\mathcal{Q}$ until the labeling budget is reached. Given the existing queried samples, the probability of drawing another query from the unlabeled set $\mathcal{U}$ is proportional to its distance to the closest sample already in the query set $\mathcal{Q}$:

$$p_{\text{query}}(\boldsymbol{x}_i) = \text{softmax}\big(h(\boldsymbol{x}_i)/\tau\big) \quad \forall i \in \mathcal{U}, \tag{2}$$

The temperature parameter τ controls the diversity of the sampling procedure, and $h(\boldsymbol{x}_i) = \min_{\boldsymbol{x}_j \in \mathcal{Q}} d(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is the distance of a sample $\boldsymbol{x}_i$ to the query set $\mathcal{Q}$. For a meaningful notion of distance, we define d in an embedding space as $d(\boldsymbol{x}, \boldsymbol{x}') = \|\phi(\boldsymbol{x}) - \phi(\boldsymbol{x}')\|_2$, where ϕ is a neural feature map. We stress that all deep methods considered in this paper have an associated feature map that we can use. The fact that L2 distance is used in the querying strategy is not an ad-hoc choice but rather aligned with the δ-ball radius definition (Eq. (5) in Supp. A) in Thm. 1.

In Supp. A, we discuss the cover radius and empirically validate that diverse querying leads to smaller δ than others

[1] This has complexity $O(KN)$ which can be reduced to $O(K \log N)$ using scalable alternatives (Bahmani et al., 2012).

and is hence advantageous for AD.

2.5. Semi-supervised Outlier Exposure Loss (SOEL)

We next consider how to use both labeled and unlabeled samples in training. We propose SOEL whose loss combines the unsupervised AD loss of LOE (Qiu et al., 2022a) for the unlabeled data with the supervised loss (Eq. (1)) for the labeled samples. For all queried data (with index set $\mathcal{Q}$), we assume that ground truth labels y_i are available, while for unqueried data (with index set $\mathcal{U}$), the labels $\tilde{y}_i$ are unknown. Adding both losses together yields

$$\mathcal{L}(\theta, \tilde{\boldsymbol{y}}) = \frac{1}{|\mathcal{Q}|} \sum_{j \in \mathcal{Q}} \big(y_j L_1^\theta(\boldsymbol{x}_j) + (1 - y_j) L_0^\theta(\boldsymbol{x}_j)\big) + \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \big(\tilde{y}_i L_1^\theta(\boldsymbol{x}_i) + (1 - \tilde{y}_i) L_0^\theta(\boldsymbol{x}_i)\big). \quad (3)$$

Similar to Qiu et al. (2022a), optimizing this loss involves a block coordinate ascent scheme that alternates between inferring the unknown labels and taking gradient steps to minimize Eq. (3) with the inferred labels. In each iteration, the pseudo labels $\tilde{y}_i$ for $i \in \mathcal{U}$ are obtained by minimizing Eq. (3) subject to a constraint of $\sum_{i \in \mathcal{Q}} y_i + \sum_{i \in \mathcal{U}} \tilde{y}_i = \alpha N$. The constraint ensures that the inferred anomaly labels respect a certain contamination ratio α. To be specific, let $\tilde{\alpha}$ denote the fraction of anomalies among the *unqueried* set $\mathcal{U}$, so that $\tilde{\alpha}|\mathcal{U}| + \sum_{j \in \mathcal{Q}} y_j = \alpha N$. The constrained optimization problem is then solved by using the current anomaly score function S to rank the unlabeled samples and assign the top $\tilde{\alpha}$-quantile of the associated labels $\tilde{y}_i$ to the value 1, and the remaining to the value 0.

We illustrate SOEL's effect on a 2D toy data example in Fig. 1, where SOEL (**d**) almost achieves the same performance as the supervised AD (**c**) with only one queried point.

In theory, α could be treated as a hyperparameter, but eliminating hyperparameters is important in AD. In many practical applications of AD, there is no labeled data that can be used for validation. While Qiu et al. (2022a) have to assume that the contamination ratio is given, SOEL provides an opportunity to estimate α. In Sec. 2.6, we develop an importance-sampling based approach to estimate α from the labeled data. Estimating this ratio can be beneficial for many AD algorithms, including OC-SVM (Schölkopf et al., 2001), kNN (Ramaswamy et al., 2000), Robust PCA/Autoencoder (Zhou and Paffenroth, 2017), and Soft-boundary deep SVDD (Ruff et al., 2018). When working with contaminated data, these algorithms require a decent estimate of the contamination ratio for good performance.

Another noteworthy aspect of the SOEL loss is that it weighs the *averaged* losses equally to each other. In Supp. E.9, we empirically show that equal weighting yields the best results among a large range of various weights. This provides more weight to every queried data point than to an unqueried one, because we expect the labeled samples to be more informative. On the other hand, it ensures that neither loss component will dominate the learning task. Our equal weighting scheme is also practical because it avoids a hyperparameter.

2.6. Contamination Ratio Estimation.

To eliminate a critical hyperparameter in our approach, we estimate the *contamination ratio* α, i.e., the fraction of anomalies in the dataset. Under a few assumptions, we show how to estimate this parameter using mini-batches composed of on non-i.i.d. samples.

We consider the contamination ratio α as the fraction of anomalies in the data. We draw on the notation from Sec. 2.1 to define $y(\boldsymbol{x})$ as an oracle, outputting 1 if $\boldsymbol{x}$ is an anomaly, and 0 otherwise (e.g., upon *querying* $\boldsymbol{x}$). We can now write $\alpha = \mathbb{E}_{p(\boldsymbol{x})}[y(\boldsymbol{x})]$.

Estimating α would be trivial given an unlimited querying budget of i.i.d. data samples. The difficulty arises due to the fact that (1) our querying budget is limited, and (2) we query data in a non-i.i.d. fashion so that the sample average is not representative of the anomaly ratio of the full data set.

Since the queried data points are not independently sampled, we cannot straightforwardly estimate α based on the empirical frequency of anomalies in the query $\mathcal{Q}$. More precisely, our querying procedure results in a chain of indices $\mathcal{Q} = \{i_1, i_2, ..., i_{|\mathcal{Q}|}\}$, where $i_1 \sim \mathrm{Unif}(1 : \mathrm{N})$, and each conditional distribution $i_k | i_{<k}$ is defined by Eq. (2). We will show as follows that this sampling bias can be compensated using importance weights.

As follows, we first propose an importance-weighted estimator of α and then prove the estimator is unbiased under certain idealized conditions specified by two assumptions about our querying strategy. Justifications for the two assumptions will be provided below.

For a random query $\mathcal{Q}$, its anomaly scores $\{S(\boldsymbol{x}_i) : i \in \mathcal{Q}\}$ and anomaly labels $\{y(\boldsymbol{x}_i) : i \in \mathcal{Q}\}$ are known. Write $S(\boldsymbol{x}_i)$ as s_i and let $p_s(s_i)$ denote the marginal density of population anomaly scores and $q_s(s_i)$ denote the marginal density of the queried samples' anomaly scores. Our importance-weighted estimator of the contamination ratio is

$$\hat{\alpha} = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \frac{p_s(s_i)}{q_s(s_i)} y(\boldsymbol{x}_i). \quad (4)$$

As discussed above, $y(\boldsymbol{x}_i)$ are the ground truth anomaly labels, obtained from querying $\mathcal{Q}$. The estimator takes into account that, upon repulsive sampling, we will sample data points in the tail regions of the data distribution more often than we would upon uniform sampling.

In practice, we learn p_s and q_s using a kernel density esti-

mator in the one-dimensional space of anomaly scores of the training data and the queried data, respectively. We set the bandwidth to the average spacing of scores. With the following two assumptions, Eq. (4) is unbiased.

Assumption 1. *The anomaly scores $\{S(\boldsymbol{x}_i) : i \in \mathcal{Q}\}$ in a query set $\mathcal{Q}$ are approximately independently distributed.*

Assumption 2. *Let $y_s(S(\boldsymbol{x}))$ denote an oracle that assigns ground truth anomaly labels based on the model's anomaly scores $S(\boldsymbol{x})$. We assume that such an oracle exists, i.e., the anomaly score $S(\boldsymbol{x})$ is a sufficient statistic of the ground truth anomaly labeling function: $y_s(S(\boldsymbol{x})) = y(\boldsymbol{x})$.*

Assumptions 1 and 2 are only approximations of reality. In our experiment section, we will show that they are good working assumptions to estimate anomaly ratios. Below, we will provide additional strong evidence that assumptions 1 and 2 are well justified.

The following theorem is a consequence of them:

Theorem 2. *Assume that Assumptions 1 and 2 hold. Then, Eq.* (4) *is an unbiased estimator of the contamination ratio α, i.e., $\mathbb{E}[\hat{\alpha}] = \alpha$.*

The proof is in Supp. B. Thm. 2 allows us to estimate the contamination ratio based on a non-iid query set $\mathcal{Q}$.

Discussion. We empirically verified the fact that Thm. 2 results in reliable estimates for varying contamination ratios in Supp. B.4. Since Assumptions 1 and 2 seem strong, we discuss their justifications and empirical validity next.

While verifying the independence assumption (Assumption 1) rigorously is difficult, we tested for linear correlations between the scores (Supp. B.2). We found that the absolute off-diagonal coefficient values are significantly smaller than one on CIFAR-10, providing support for Assumption 1. A heuristic argument can be provided to support the validity of Assumption 1 based on the following intuition. When data points are sampled diversely in a high-dimensional space, the negative correlations induced by their repulsive nature tend to diminish when the data is projected onto a one-dimensional subspace. This intuition stems from the fact that a high-dimensional ambient space offers ample dimensions for the data points to avoid clustering. To illustrate this, consider the scenario of sampling diverse locations on the Earth's surface, with each location representing a point in the high-dimensional space. By including points from various continents, we ensure diversity in their spatial distribution. However, when focusing solely on the altitude of these locations (such as distinguishing between mountain tops and flat land), it is plausible that the altitude levels are completely uncorrelated. While this heuristic argument provides an intuitive understanding, it is important to note that it does not offer a rigorous mathematical proof.

To test Assumption 2, we tested the degree to which the anomaly score is a sufficient statistic for anomaly scoring on the training set. The assumption would be violated if we could find pairs of training data $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where $\boldsymbol{x}_i \neq \boldsymbol{x}_j$, with identical anomaly scores $S(\boldsymbol{x}_i) = S(\boldsymbol{x}_j)$ but different anomaly labels $y_s(s_i) \neq y_s(s_j)$[2]. On FMNIST, we found 38 data pairs with matching scores, and none of them had opposite anomaly labels. For CIFAR-10, the numbers were 21 and 3, respectively. See Supp. B.3 for details.

3. Related Work

Deep Anomaly Detection. Many recent advances in anomaly detection are in the area of deep learning (Ruff et al., 2021). One early strategy was to use autoencoder- (Zhou and Paffenroth, 2017; Principi et al., 2017) or density-based models (Schlegl et al., 2017; Deecke et al., 2018). Another pioneering stream of research combines one-class classification (Schölkopf et al., 2001) with deep learning for unsupervised (Ruff et al., 2018; Qiu et al., 2022b) and semi-supervised (Ruff et al., 2019) anomaly detection. Many other approaches to deep anomaly detection are self-supervised. They employ a self-supervised loss function to train the detector and score anomalies (Qiu et al., 2021; Golan and El-Yaniv, 2018; Hendrycks et al., 2019; Bergman and Hoshen, 2020; Shenkar and Wolf, 2022; Schneider et al., 2022). Our work resides in the self-supervised anomaly detection category and can be extended to other data modalities if an appropriate loss is provided.

While all these methods assume that the training data consists of only normal samples, in many practical applications, the training pool may be contaminated with unidentified anomalies (Vilhjálmsson and Nordborg, 2013; Steinhardt et al., 2017). This can be problematic because the detection accuracy typically deteriorates when the contamination ratio increases (Wang et al., 2019). Addressing this, refinement (Zhou and Paffenroth, 2017; Yoon et al., 2021) attempts to cleanse the training pool by removing anomalies therein, although they may provide valuable training signals. As a remedy, Qiu et al. (2022a) propose to jointly infer binary labels to each datum (normal vs. anomalous) while updating the model parameters based on outlier exposure. Our work also makes the contaminated data assumption and employs the training signal of abnormal data.

Querying Strategies for Anomaly Detection. Querying strategies play an important role in batch active learning (Sener and Savarese, 2018; Ash et al., 2020; Citovsky et al., 2021; Pinsler et al., 2019; Hoi et al., 2006) but are less studied for anomaly detection. The human-in-the-loop setup for anomaly detection has been pioneered by Pelleg and Moore (2004). Query samples are typically chosen locally, e.g.,

[2]The condition $S(\boldsymbol{x}_i) \neq S(\boldsymbol{x}_j)$ for $\boldsymbol{x}_i \neq \boldsymbol{x}_j$ hints we can assign a unique label to each data point based on their scores.

close to the decision boundary of a one-class SVM (Görnitz et al., 2013; Yin et al., 2018) or sampled according to a density model (Ghasemi et al., 2011). Siddiqui et al. (2018); Das et al. (2016) propose to query the most anomalous instance, while Das et al. (2019) employ a tree-based ensemble to query both anomalous and diverse samples. A recent survey compares various aforementioned query strategies with one-class classifiers (Trittenbach et al., 2021).

Pimentel et al. (2020) query samples with the top anomaly scores for autoencoder-based methods, while Ning et al. (2022) improve the querying by considering the diversity. Tang et al. (2020) use an ensemble of deep anomaly detectors and query the most likely anomalies for each detector separately. Russo et al. (2020) query samples where the model is uncertain about the predictions. Pang et al. (2021) and Zha et al. (2020) propose querying strategies based on reinforcement learning, which requires labeled datasets.

All these querying strategies do not optimize coverage as defined in Thm. 1, and as a result, their generalization guarantees are less favorable than our method. Most querying strategies from the papers discussed above are fairly general and can be applied in combination with various backbone models. Since more powerful backbone models have been released since these earlier publications, we ensure a fair comparison by studying all querying strategies in combination with the same backbone models as SOEL.

4. Experiments

We study SOEL on standard image benchmarks, medical images, tabular data, and surveillance videos. Our extensive empirical study establishes how our proposed method compares to eight AD methods with labeling budgets implemented as baselines. We first describe the baselines and their implementations (Tab. 1) and then the experiments on images (Sec. 4.1), tabular data (Sec. 4.2), videos (Sec. 4.3) and finally additional experiments (Sec. 4.4).

Baselines. Most existing baselines apply their proposed querying and training strategies to shallow AD methods or sub-optimal deep models (e.g., autoencoders (Zhou and Paffenroth, 2017)). In recent years, these approaches have consistently been outperformed by self-supervised AD methods (Hendrycks et al., 2019). For a fair comparison, we endow all baselines with the same self-supervised backbone models also used in our method. By default we use neural transformation learning (NTL) (Qiu et al., 2021) as the backbone model, which was identified as state-of-the-art in a recent independent comparison of 13 models (Alvarez et al., 2022). Results with other backbone models are shown in Supp. E.2.

The baselines are summarized in Tab. 1 and detailed in Supp. C. They differ in their querying strategies (col. 3) and training strategies (col. 4 & 5): the unlabeled data is either ignored or modeled with a one-class objective. Most baselines incorporate the labeled data by a supervised loss (Eq. (1)). As an exception, Ning et al. (2022) remove all queried anomalies and then train a weighted one-class objective on the remaining data. All baselines weigh the unsupervised and supervised losses equally. They differ in their querying strategies, summarized below:

- **Margin query** selects samples close to the boundary of the normality region deterministically. The method uses the true contamination ratio to choose an ideal boundary.
- **Margin diverse query** combines margin query with neighborhood-based diversification. It selects samples that are not k-nearest neighbors of the queried set. Thus samples are both diverse and close to the boundary.
- **Most positive query** always selects the top-ranked samples ordered by their anomaly scores.
- **Positive diverse query** combines querying according to anomaly scores with distance-based diversification. The selection criterion combines anomaly score and the minimum Euclidean distance to all queried samples.
- **Random query** draws samples uniformly.
- **Positive random query** samples uniformly among the top 50% data ranked by anomaly scores.

Implementation Details. In all experiments, we use a NTL (Qiu et al., 2021) backbone model for all methods. Experiments with other backbone models are shown in Supp. E.2. On images and videos, NTL is built upon the penultimate layer output of a frozen ResNet-152 pre-trained on ImageNet. NTL is trained for one epoch, after which all $|\mathcal{Q}|$ queries are labeled at once. The contamination ratio α in SOEL is estimated immediately after the querying step and then fixed for the remaining training process. We follow Qiu et al. (2022a) and set $\tilde{y}_i = 0.5$ for inferred anomalies. This accounts for the uncertainty of whether the sample truly is an anomaly. More details are given in Supp. D and Alg. 1.

4.1. Experiments on Image Data

We study SOEL on standard image benchmarks to establish how it compares to eight well-known baselines with various querying and training strategies. Informative querying plays an important role in medical domains where expert labeling is expensive. Hence, we also study nine medical datasets from Yang et al. (2021). We describe the datasets, the evaluation protocol, and finally the results of our study.

Image Benchmarks. We experiment with two popular image benchmarks: CIFAR-10 and Fashion-MNIST. These have been widely used in previous papers on deep AD (Ruff et al., 2018; Golan and El-Yaniv, 2018; Hendrycks et al., 2019; Bergman and Hoshen, 2020).

Medical Images. Since medical imaging is an important practical application of AD, we also study SOEL on medi-

Table 1. A summary of all compared experimental methods' query strategy and training strategy irrespective of their backbone models.

Name	Reference	Querying Strategy	Loss (labeled)	Loss (unlabeled)
Mar	Görnitz et al. (2013)	margin query	superv. (Eq. (1))	one class
Hybr1	Görnitz et al. (2013)	margin diverse query	superv. (Eq. (1))	one class
Pos1	Pimentel et al. (2020)	most positive query	superv. (Eq. (1))	none
Pos2	Barnabé-Lortie et al. (2015)	most positive query	superv. (Eq. (1))	one class
Rand1	Ruff et al. (2019)	random query	superv. (Eq. (1))	one class
Rand2	Trittenbach et al. (2021)	positive random query	superv. (Eq. (1))	one class
Hybr2	Das et al. (2019)	positive diverse query	superv. (Eq. (1))	none
Hybr3	Ning et al. (2022)	positive diverse query	refinement	weighted one class
SOEL	[ours]	diverse (Eq. (2))	semi-supervised outlier exposure loss (Eq. (3))	

Table 2. AUC (%) with standard deviation for anomaly detection on 11 image datasets when the query budget $|\mathcal{Q}| = 20$. SOEL outperforms all baselines by a large margin by querying diverse and informative samples.

	Mar	**Hybr1**	**Pos1**	**Pos2**	**Rand1**	**Rand2**	**Hybr2**	**Hybr3**	**SOEL**
CIFAR10	92.4±0.7	92.0±0.7	93.4±0.5	92.1±0.7	89.2±3.2	91.4±1.0	85.1±2.2	71.8±7.4	**96.3±0.3**
FMNIST	93.1±0.4	92.6±0.4	92.2±0.6	89.3±1.0	84.0±3.8	90.6±1.1	88.7±1.4	82.6±4.3	**94.8±0.6**
Blood	68.6±1.8	69.1±1.3	69.6±1.8	72.2±4.9	70.6±1.6	69.2±1.7	72.2±2.7	58.3±5.2	**80.5±0.5**
OrganA	86.4±1.3	87.4±0.7	81.7±2.9	81.8±2.1	82.9±0.6	86.5±0.7	88.6±1.5	68.8±3.1	**90.7±0.7**
OrganC	86.5±0.9	87.0±0.7	84.6±1.9	79.6±2.0	85.5±0.9	86.4±0.8	84.8±1.2	68.9±3.0	**89.7±0.7**
OrganS	83.5±1.1	84.1±0.4	83.2±1.3	78.6±1.0	82.2±1.4	83.8±0.4	82.3±0.7	66.9±4.3	**87.4±0.8**
OCT	64.4±3.7	63.3±1.8	63.8±4.4	63.0±4.0	59.7±1.9	62.1±4.3	63.0±7.6	56.2±4.5	**68.5±3.4**
Path	82.7±2.4	86.0±1.1	77.5±2.0	80.2±3.5	83.2±1.6	83.9±2.9	86.1±2.0	75.1±4.2	**88.1±1.1**
Pneumonia	72.1±7.0	75.1±5.3	75.5±8.8	83.6±6.1	68.1±5.9	76.0±8.0	88.4±3.3	63.4±17.7	**91.2±1.4**
Tissue	60.2±1.5	61.3±1.7	65.8±1.7	63.5±2.0	59.9±1.7	59.5±1.3	62.1±1.7	50.8±1.6	**66.4±1.4**
Derma	62.6±3.8	63.1±4.7	66.6±2.3	66.4±4.3	64.5±4.8	68.3±2.1	57.2±13.3	48.0±13.6	**73.5±2.5**
Average	77.5	78.3	77.3	77.6	75.4	78.0	78.0	64.6	**84.3**

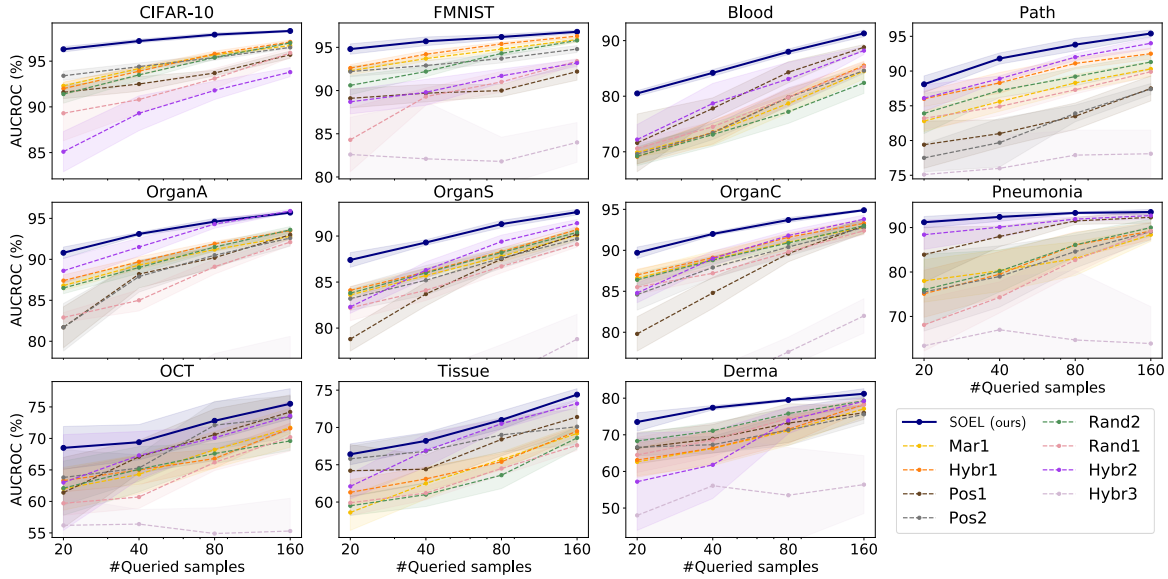

Figure 2. Running AUCs (%) with different query budgets. Models are evaluated at $20, 40, 80, 160$ queries. SOEL performs the best among the compared methods on all query budgets.

cal images. The datasets we consider cover different data modalities (e.g., X-ray, CT, electron microscope) and their characteristic image features can be very different from natural images. Our empirical study includes all 2D image datasets presented in Yang et al. (2021) that have more than 500 samples in each class, including Blood, OrganA, OrganC, OrganS, OCT, Pathology, Pneumonia, and Tissue. We also include Dermatoscope but restricted to the classes with more than 500 training samples.

Evaluation Protocol. We follow the community standard known as the "one-vs.-rest" protocol to turn these classification datasets into a test-bed for AD (Ruff et al., 2018;

Hendrycks et al., 2019; Bergman and Hoshen, 2020). While respecting the original train and test split of these datasets, the protocol iterates over the classes and treats each class in turn as normal. Random samples from the other classes are used to contaminate the data. The training set is then a mixture of unlabeled normal and abnormal samples with a contamination ratio of 10% (Wang et al., 2019; Qiu et al., 2022a; Ruff et al., 2019). This protocol can simulate a "human expert" to provide labels for the queried samples because the datasets provide ground-truth class labels. The reported results (in terms of AUC $\%$) for each dataset are averaged over the number of experiments (i.e., classes) and over five independent runs.

Results. We report the evaluation results of our method (SOEL) and the eight baselines on all eleven image datasets in Tab. 2. When querying 20 samples, our proposed method SOEL significantly outperforms the best-performing baseline by 6 percentage points on average across all datasets. We also study detection performance as the query budget increases from 20 to 160 in Fig. 2. The results show that, with a small budget of 20 samples, SOEL (by querying diverse and informative samples) makes better usage of the labels than the other baselines and thus leads to better performance by a large margin. As more samples are queried, the performance of almost all methods increases but even for 160 queries when the added benefit from adding more queries starts to saturate, SOEL still outperforms the baselines.

4.2. Experiments on Tabular Data

Many practical use cases of AD (e.g., in health care or cyber security) are concerned with tabular data. For this reason, we study SOEL on four tabular datasets from various domains. We find that it outperforms existing baselines, even with as few as 10 queries. We also confirmed the fact that our deep models are competitive with classical methods for tabular data in Supp. E.13.

Tabular Datasets. Our study includes the four multi-dimensional tabular datasets from the ODDS repository which have an outlier ratio of at least 30%. This is necessary to ensure that there are enough anomalies available to remove from the test set and add to the clean training set (which is randomly sub-sampled to half its size) to achieve a contamination ratio of 10%. The datasets are BreastW, Ionosphere, Pima, and Satellite. As in the image experiments, there is one round of querying, in which 10 samples are labeled. For each dataset, we report the averaged F1-score ($\%$) with standard deviations over five runs with random train-test splits and random initialization.

Results. SOEL performs best on 3 of 4 datasets and outperforms all baselines by 3.2 percentage points on average. Diverse querying best utilizes the query budget to label the diverse and informative data points, yielding a consistent improvement over existing baselines on tabular data.

4.3. Experiments on Video Data

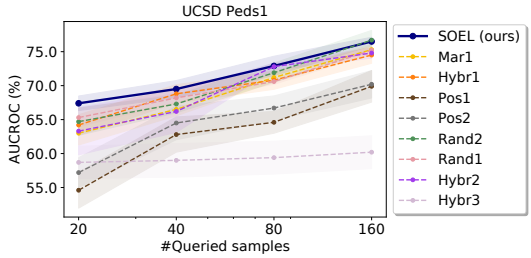

Figure 3. Results on the video dataset UCSD Peds1 with different query budgets. SOEL achieves the leading performance.

Detecting unusual objects in surveillance videos is an important application area for AD. Due to the large variability in abnormal objects and suspicious behavior in surveillance videos, expert feedback is very valuable to train an anomaly detector in a semi-supervised manner. We use NTL as the backbone model and study SOEL on a public surveillance video dataset (UCSD Peds1). The goal is to detect abnormal video frames that contain non-pedestrian objects.

Following Pang et al. (2020), we subsample a mix of normal and abnormal frames for training (using an anomaly ratio of 0.3) and use the remaining frames for testing. Before running any of the methods, a ResNet pretrained on ImageNet is used to obtain a fixed feature vector for each frame. We vary the query budget from $|\mathcal{Q}| = 20$ to $|\mathcal{Q}| = 160$ and compare SOEL to all baselines. Results in terms of average AUC and standard error over five independent runs are reported in Fig. 3. SOEL consistently outperforms all baselines, especially for smaller querying budgets.

4.4. Additional Experiments

In Supp. E, we provide additional experiments and ablations demonstrating SOEL's strong performance and justifying modeling choices. The three most important findings are:

- **SOEL vs. Active Learning:** Our framework is superior to its extension to the sequential active learning (Fig. 9).
- **Varying Contamination Ratio:** Fig. 7 demonstrates that SOEL dominates under varying contamination ratios (1%, 5%, 20%). In addition, Tab. 4 confirms that Eq. (4) reliably estimates α on both CIFAR-10 and F-MNIST.
- **Backbone Models:** Tab. 6 shows that SOEL also performs best for the backbone models MHRot (Hendrycks et al., 2019) and DSVDD (Ruff et al., 2018).

In addition, we provide an ablation of the temperature τ (Tab. 8), a discussion on the effects of initialization randomness (Supp. E.1), an ablation study of the pseudo-label $\tilde{y}$ values (Tab. 9), a comparison to binary classification (Fig. 8), an ablation of the SOEL loss components, an ablation of querying strategies (Fig. 10), additional methods

Table 3. F1-score (%) with standard deviation for anomaly detection on tabular data when the query budget $|\mathcal{Q}| = 10$. SOEL performs the best on 3 of 4 datasets and outperforms all baselines by 3.2 percentage points on average.

	Mar	Hybr1	Pos1	Pos2	Rand1	Rand2	Hybr2	Hybr3	SOEL
BreastW	81.6±0.7	83.3±2.0	58.6±7.7	81.3±0.8	87.1±1.0	82.9±1.1	55.0±6.0	79.6±4.9	**93.9±0.5**
Ionosphere	91.9±0.3	**92.3±0.5**	56.1±6.2	91.1±0.8	91.1±0.3	91.9±0.6	64.0±4.6	88.2±0.9	91.8±1.1
Pima	50.1±1.3	49.2±1.9	48.5±0.4	52.4±0.8	53.6±1.1	51.9±2.0	53.8±4.0	48.4±0.7	**55.5±1.2**
Satellite	64.2±1.2	66.2±1.7	57.0±3.0	56.7±3.2	67.7±1.2	66.6±0.8	48.6±6.9	56.9±7.0	**71.1±1.7**
Average	72.0	72.8	55.1	70.4	74.9	73.3	55.4	68.3	**78.1**

for inferring y (Fig. 13), and comparison to additional semi-supervised baselines (Fig. 13, Tab. 10).

5. Conclusion

We introduced semi-supervised outlier exposure with a limited labeling budget (SOEL). Inspired by a set of conditions that guarantee the generalization of anomaly score rankings from queried to unqueried data, we proposed to use a diversified querying strategy and a combination of two losses for queried and unqueried samples. By weighting the losses equally to each other and by estimating the unknown contamination rate from queried samples, we were able to make our approach free of its most important hyperparameters, making it easy to use. An extensive empirical study on images, tabular data, and video confirmed the efficacy of SOEL as a semi-supervised learning framework compatible with many existing losses for AD.

Limitations: The success of our approach relies on several heuristics that we demonstrated were empirically effective but that cannot be proven rigorously. Estimation of the contamination ratio can be noisy when the query set is small—but the LOE loss is robust even under misspecification of the contamination ratio (Qiu et al., 2022a). The diversified sampling strategy becomes expensive when the dataset is large, but this can be mitigated by random data thinning.

Societal Impacts: The use of human labels for anomaly detection runs the risk of introducing potential human biases in the definition of what is anomalous, particularly for datasets involving human subjects. Since our approach relies heavily on a relatively small number of human labels, the deployment of our approach with real human labelers would benefit by having guidelines for the labelers in terms of providing fair labels and avoiding amplification of bias.

Acknowledgements

SM acknowledges support by the National Science Foundation (NSF) under an NSF CAREER Award, award numbers 2003237 and 2007719, by the Department of Energy under grant DE-SC0022331, by the HPI Research Center in Machine Learning and Data Science at UC Irvine, by the IARPA WRIVA program, and by gifts from Qualcomm and Disney. Part of this work was conducted within the DFG research unit FOR 5359 on Deep Learning on Sparse Chemical Process Data. MK acknowledges support by the Carl-Zeiss Foundation, the DFG awards KL 2698/2-1, KL 2698/5-1, KL 2698/6-1, and KL 2698/7-1, and the BMBF awards 03|B0770E and 01|S21010C.

The Bosch Group is carbon neutral. Administration, manufacturing and research activities do no longer leave a carbon footprint. This also includes GPU clusters on which the experiments have been performed.

References

Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 2021.

Siqi Wang, Yijie Zeng, Xinwang Liu, En Zhu, Jianping Yin, Chuanfu Xu, and Marius Kloft. Effective end-to-end unsupervised outlier detection via inlier priority of discriminative network. In *Advances in Neural Information Processing Systems*, pages 5962–5975, 2019.

Chen Qiu, Aodong Li, Marius Kloft, Maja Rudolph, and Stephan Mandt. Latent outlier exposure for anomaly detection with contaminated data. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 18153–18167. PMLR, 2022a.

Holger Trittenbach, Adrian Englhardt, and Klemens Böhm. An overview and a benchmark of active learning for outlier detection with one-class classifiers. *Expert Systems with Applications*, 168:114372, 2021.

Chong Zhou and Randy C Paffenroth. Anomaly detection with robust deep autoencoders. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 665–674, 2017.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pages 4393–4402. PMLR, 2018.

Chen Qiu, Timo Pfrommer, Marius Kloft, Stephan Mandt, and Maja Rudolph. Neural transformation learning for deep anomaly detection beyond images. In *International Conference on Machine Learning*, pages 8703–8714. PMLR, 2021.

Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

Nico Görnitz, Marius Kloft, Konrad Rieck, and Ulf Brefeld. Toward supervised anomaly detection. *Journal of Artificial Intelligence Research*, 46:235–262, 2013.

Bahman Bahmani, Benjamin Moseley, Andrea Vattani, Ravi Kumar, and Sergei Vassilvitskii. Scalable k-means+. *Proceedings of the VLDB Endowment*, 5(7), 2012.

Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

Sridhar Ramaswamy, Rajeev Rastogi, and Kyuseok Shim. Efficient algorithms for mining outliers from large data sets. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 427–438, 2000.

Emanuele Principi, Fabio Vesperini, Stefano Squartini, and Francesco Piazza. Acoustic novelty detection with adversarial autoencoders. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3324–3330. IEEE, 2017.

Thomas Schlegl, Philipp Seeböck, Sebastian M Waldstein, Ursula Schmidt-Erfurth, and Georg Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In *International conference on information processing in medical imaging*, pages 146–157. Springer, 2017.

Lucas Deecke, Robert Vandermeulen, Lukas Ruff, Stephan Mandt, and Marius Kloft. Image anomaly detection with generative adversarial networks. In *Joint european conference on machine learning and knowledge discovery in databases*, pages 3–17. Springer, 2018.

Chen Qiu, Marius Kloft, Stephan Mandt, and Maja Rudolph. Raising the bar in graph-level anomaly detection. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 2196–2203, 2022b.

Izhak Golan and Ran El-Yaniv. Deep anomaly detection using geometric transformations. In *Advances in Neural Information Processing Systems*, pages 9758–9769, 2018.

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in Neural Information Processing Systems*, 32:15663–15674, 2019.

Liron Bergman and Yedid Hoshen. Classification-based anomaly detection for general data. In *International Conference on Learning Representations*, 2020.

Tom Shenkar and Lior Wolf. Anomaly detection for tabular data with internal contrastive learning. In *International Conference on Learning Representations*, 2022.

Tim Schneider, Chen Qiu, Marius Kloft, Decky Aspandi Latif, Steffen Staab, Stephan Mandt, and Maja Rudolph. Detecting anomalies within time series using local neural transformations. *arXiv preprint arXiv:2202.03944*, 2022.

Bjarni J Vilhjálmsson and Magnus Nordborg. The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, 14(1):1–2, 2013.

Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3520–3532, 2017.

Jinsung Yoon, Kihyuk Sohn, Chun-Liang Li, Sercan O Arik, Chen-Yu Lee, and Tomas Pfister. Self-trained one-class classification for unsupervised anomaly detection. *arXiv preprint arXiv:2106.06115*, 2021.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018.

Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020.

Gui Citovsky, Giulia DeSalvo, Claudio Gentile, Lazaros Karydas, Anand Rajagopalan, Afshin Rostamizadeh, and Sanjiv Kumar. Batch active learning at scale. *Advances in Neural Information Processing Systems*, 34:11933–11944, 2021.

Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. *Advances in neural information processing systems*, 32, 2019.

Steven CH Hoi, Rong Jin, and Michael R Lyu. Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on World Wide Web*, pages 633–642, 2006.

Dan Pelleg and Andrew Moore. Active learning for anomaly and rare-category detection. *Advances in neural information processing systems*, 17, 2004.

Lili Yin, Huangang Wang, and Wenhui Fan. Active learning based support vector data description method for robust novelty detection. *Knowledge-Based Systems*, 153:40–52, 2018.

Alireza Ghasemi, Hamid R Rabiee, Mohsen Fadaee, Mohammad T Manzuri, and Mohammad H Rohban. Active learning from positive and unlabeled data. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 244–250. IEEE, 2011.

Md Amran Siddiqui, Alan Fern, Thomas G Dietterich, Ryan Wright, Alec Theriault, and David W Archer. Feedback-guided anomaly discovery via online optimization. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2200–2209, 2018.

Shubhomoy Das, Weng-Keen Wong, Thomas Dietterich, Alan Fern, and Andrew Emmott. Incorporating expert feedback into active anomaly discovery. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 853–858. IEEE, 2016.

Shubhomoy Das, Md Rakibul Islam, Nitthilan Kannappan Jayakodi, and Janardhan Rao Doppa. Active anomaly detection via ensembles: Insights, algorithms, and interpretability. *arXiv preprint arXiv:1901.08930*, 2019.

Tiago Pimentel, Marianne Monteiro, Adriano Veloso, and Nivio Ziviani. Deep active learning for anomaly detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

Jin Ning, Leiting Chen, Chuan Zhou, and Yang Wen. Deep active autoencoders for outlier detection. *Neural Processing Letters*, pages 1–13, 2022.

Xuning Tang, Yihua Shi Astle, and Craig Freeman. Deep anomaly detection with ensemble-based active learning. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1663–1670. IEEE, 2020.

Stefania Russo, Moritz Lürig, Wenjin Hao, Blake Matthews, and Kris Villez. Active learning for anomaly detection in environmental data. *Environmental Modelling & Software*, 134:104869, 2020.

Guansong Pang, Anton van den Hengel, Chunhua Shen, and Longbing Cao. Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1298–1308, 2021.

Daochen Zha, Kwei-Herng Lai, Mingyang Wan, and Xia Hu. Meta-aad: Active anomaly detection with deep reinforcement learning. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 771–780. IEEE, 2020.

Vincent Barnabé-Lortie, Colin Bellinger, and Nathalie Japkowicz. Active learning for one-class classification. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*, pages 390–395. IEEE, 2015.

Maxime Alvarez, Jean-Charles Verdier, D'Jeff K Nkashama, Marc Frappier, Pierre-Martin Tardif, and Froduald Kabanza. A revealing large-scale evaluation of unsupervised anomaly detection algorithms. *arXiv preprint arXiv:2204.09825*, 2022.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *arXiv preprint arXiv:2110.14795*, 2021.

Guansong Pang, Cheng Yan, Chunhua Shen, Anton van den Hengel, and Xiao Bai. Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12173–12182, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16, 2003.

Andreas Krause and Daniel Golovin. Submodular function maximization. *Tractability*, 3:71–104, 2014.

David Arthur and Sergei Vassilvitskii. K-means++ the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035, 2007.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020a.

Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019.

Zhun Li, ByungSoo Ko, and HoJin Choi. Pseudo-labeling using gaussian process for semi-supervised deep learning. In *2018 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pages 263–269. IEEE, 2018.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

David M. J. Tax and Robert P. W. Duin. Support vector data description. *Mach. Learn.*, 54(1):45–66, 2004.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE, 2008.

Max Welling Diederik P. Kingma. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Alireza Makhzani and Brendan J Frey. Winner-take-all autoencoders. *Advances in neural information processing systems*, 28, 2015.

Kihyuk Sohn, Chun-Liang Li, Jinsung Yoon, Minho Jin, and Tomas Pfister. Learning and evaluating representations for deep one-class classification. *arXiv preprint arXiv:2011.02578*, 2020b.

A. Theorem 1

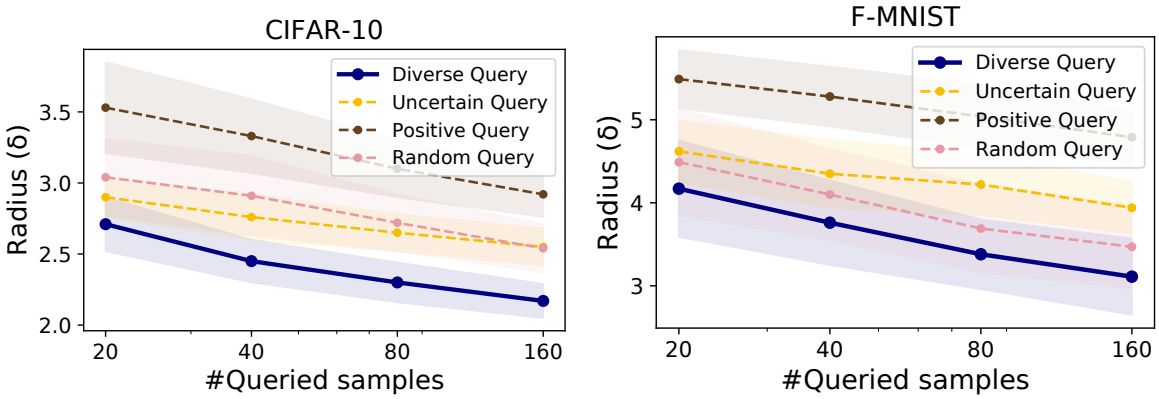

Figure 4. Cover radius δ (Eq. (5)) resulted from different querying strategies on the first class of CIFAR-10 and F-MNIST. Diverse queries systematically have smaller cover radius than other querying strategies.

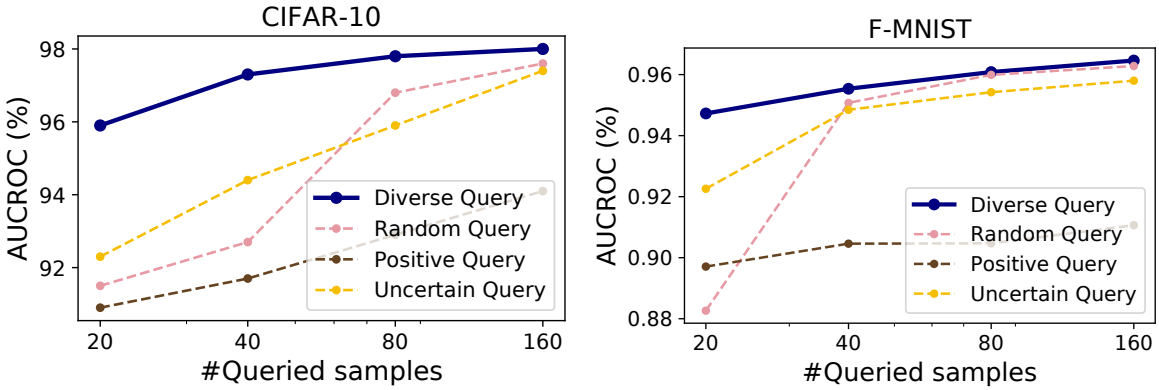

Figure 5. Ranking performance of unlabeled data. AUC of unqueried data is evaluated using the fitted anomaly detector on the queried data. Our proposed diverse querying (k-means++) provides better ranking of the unlabeled data.

Proof. Since S is λ_s-Lipschitz continuous and $\boldsymbol{u}_a$ and $\boldsymbol{u}_n$ are assumed to be closer than δ to $\boldsymbol{x}_a$ and $\boldsymbol{x}_n$ respectively, we have $S(\boldsymbol{x}_a) - \delta\lambda_s \leq S(\boldsymbol{u}_a)$ and $-S(\boldsymbol{x}_n) - \delta\lambda_s \leq -S(\boldsymbol{u}_n)$. Adding the inequalities and using the condition $S(\boldsymbol{x}_a) - S(\boldsymbol{x}_n) \geq 2\delta\lambda_s$, yields $0 \leq S(\boldsymbol{x}_a) - S(\boldsymbol{x}_n) - 2\delta\lambda_s \leq S(\boldsymbol{u}_a) - S(\boldsymbol{u}_n)$, which proves $S(\boldsymbol{u}_a) \geq S(\boldsymbol{u}_n)$. □

In Thm. 1, we considered using the fixed-radius neighborhood (δ-ball) of the queried data as the cover of the whole dataset, and mentioned diverse querying has a smaller radius than other querying strategies. In this section, we will empirically verify this fact and further illustrate diverse querying leads to good ranking of un-queried data (see also Fig. 10 on test data).

As defined in Thm. 1, the radius is the smallest distance that is required for any un-queried sample to be covered by the neighborhood of a queried sample of the same type. Mathematically, we compute the radius as

$$\delta = \max_{i \in \mathcal{U}} \min_{j \in \mathcal{Q}, y_i = y_j} d(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{5}$$

where we adopt the euclidean distance in the feature space for a meaningful metric d. We apply NTL on the first class of CIFAR-10 and F-MNIST dataset. We make queries with different budgets, after which we compute δ by Eq. (5). We repeat this procedure for 100 times and report the mean and standard deviation in Fig. 4. We compared four querying strategies: diverse queries (k-means++), uncertain queries (Mar), positive queries (Pos1), and random queries (Rand1). It shows that diverse queries significantly lead the smallest radius δ among the compared strategies on all querying budgets.

Next, we provide an empirical, overall justification of Thm. 1 (see also Fig. 10 on test data). An implication of Thm. 1 is that, assuming anomaly scores are fixed, a smaller δ will satisfy the large anomaly score margin ($S(\boldsymbol{x}_a) - S(\boldsymbol{x}_n)$) more easily, hence it is easier for S to correctly rank the remaining unlabeled points. To justify this implication, we need a metric of ranking. AUC satisfies this requirement as it is alternatively defined as (Mohri et al., 2018, 10.5.2)(Cortes and Mohri, 2003)

$$\text{AUC} = \frac{1}{|\mathcal{U}_0| + |\mathcal{U}_1|} \sum_{n \in \mathcal{U}_0, a \in \mathcal{U}_1} \mathbb{1}(S(\boldsymbol{u}_a) > S(\boldsymbol{u}_n)) \approx P_{n \in \mathcal{U}_0, a \in \mathcal{U}_1}(S(\boldsymbol{u}_a) > S(\boldsymbol{u}_n))$$

which measures the probability of ranking unlabeled samples $\boldsymbol{u}_a$ higher than $\boldsymbol{u}_n$ in terms of their scores. $\mathcal{U} = \mathcal{U}_0 \bigcup \mathcal{U}_1$ is the un-queried data indices and $\mathcal{U}_0$ and $\mathcal{U}_1$ are disjoint un-queried normal and abnormal data sets respectively. $\boldsymbol{u}_a$ and $\boldsymbol{u}_n$ are instances of each kind. We conducted experiments on CIFAR-10 and F-MNIST, where we trained an anomaly detector (NTL) on the queried data for 30 epochs and then compute the AUC on the remaining un-queried data. The results of four querying straties are reported in Fig. 5, which shows that our proposed diverse querying strategy generalizes the anomaly score ranking the best to the unqueried data among the compared strategies, testifying our analysis in the main paper. A consequence is that diverse querying can provide accurate assignments of the latent anomaly labels, which will further help learn a high-quality of anomaly detector through the unsupervised loss term in Eq. (3).

Optimality of Cover Radius. Although k-means++ greedily samples the queries which may have a sub-optimal cover radius, greedy sampling strategies for selecting a diverse set of datapoints in a multi-dimensional space are known to produce nearly optimal solutions (Krause and Golovin, 2014), with significant runtime savings over more sophisticated search methods. As a results, we follow common practice (e.g. Arthur and Vassilvitskii (2007)) and also use the greedy approach. We check the diversity of the rustling query set by comparing all sampling strategies considered in the paper in terms of data coverage. Figure 4 shows that the greedy strategy we use achieves the best coverage, i.e. results in the most diverse query set.

On the Assumptions of Thm. 1. In the proof, we assume a Lipschitz continuous S and a large margin between $S(\boldsymbol{x}_a)$ and $S(\boldsymbol{x}_n)$. Lipschitz continuity serves as a working assumption and is a common assumption when analyzing optimization landscapes of deep learning. Lipschitz continuity can be controlled by the strength of regularization on the model parameters. The large margin condition is achieved by optimizing our loss function. The supervised anomaly detection loss encourages a large margin as it minimizes the anomaly score of queried normal data and maximizes the score of the queried abnormal data. If the anomaly score function doesn't do well for the queried samples, then it should be optimized further. Our empirical results also show this is a reasonable condition.

B. Theorem 2

In this section, we will empirically justify the assumptions we made in Sec. 2.6 that are used to build an unbiased estimator of the anomaly ratio α (Eq. (4)). We will also demonstrate the robustness of the estimation under varying α.

B.1. Proof

Proof. Let A1 and A2 denote Assumption 1 and 2, respectively. Furthermore, let $q(\boldsymbol{x}_1, ..., \boldsymbol{x}_{|\mathcal{Q}|})$ and $q_s(s_1, ..., s_{|\mathcal{Q}|})$ denote the query distribution in the data and anomaly score spaces, respectively. A2 assumes $y_s(s) := y_s(S(\boldsymbol{x})) = y(\boldsymbol{x})$ for all $\boldsymbol{x}$. So the expectation of Eq. (4) is

$$\begin{aligned}
\mathbb{E}[\hat{\alpha}] &= \mathbb{E}_{q(\boldsymbol{x}_1,...,\boldsymbol{x}_{|\mathcal{Q}|})} \left[\frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \frac{p_s(S(\boldsymbol{x}_i))}{q_s(S(\boldsymbol{x}_i))} y(\boldsymbol{x}) \right] \overset{A2}{=} \mathbb{E}_{q_s(s_1,...,s_{|\mathcal{Q}|})} \left[\frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \frac{p_s(s_i)}{q_s(s_i)} y_s(s_i) \right] \\
&\overset{A1}{=} \mathbb{E}_{\prod_{i=1}^{|\mathcal{Q}|} q_s(s_i)} \left[\frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \frac{p_s(s_i)}{q_s(s_i)} y_s(s_i) \right] = \frac{1}{|\mathcal{Q}|} \sum_{i=1}^{|\mathcal{Q}|} \mathbb{E}_{q_s(s_i)} \left[\frac{p_s(s_i)}{q_s(s_i)} y_s(s_i) \right] = \mathbb{E}_{p_s(s)}[y_s(s)] \\
&= \mathbb{E}_{p(\boldsymbol{x})}[y_s(S(\boldsymbol{x}))] \overset{A2}{=} \mathbb{E}_{p(\boldsymbol{x})}[y(\boldsymbol{x})] = \alpha
\end{aligned}$$

where the change of variables makes necessary assumptions, including the existence of density functions. □

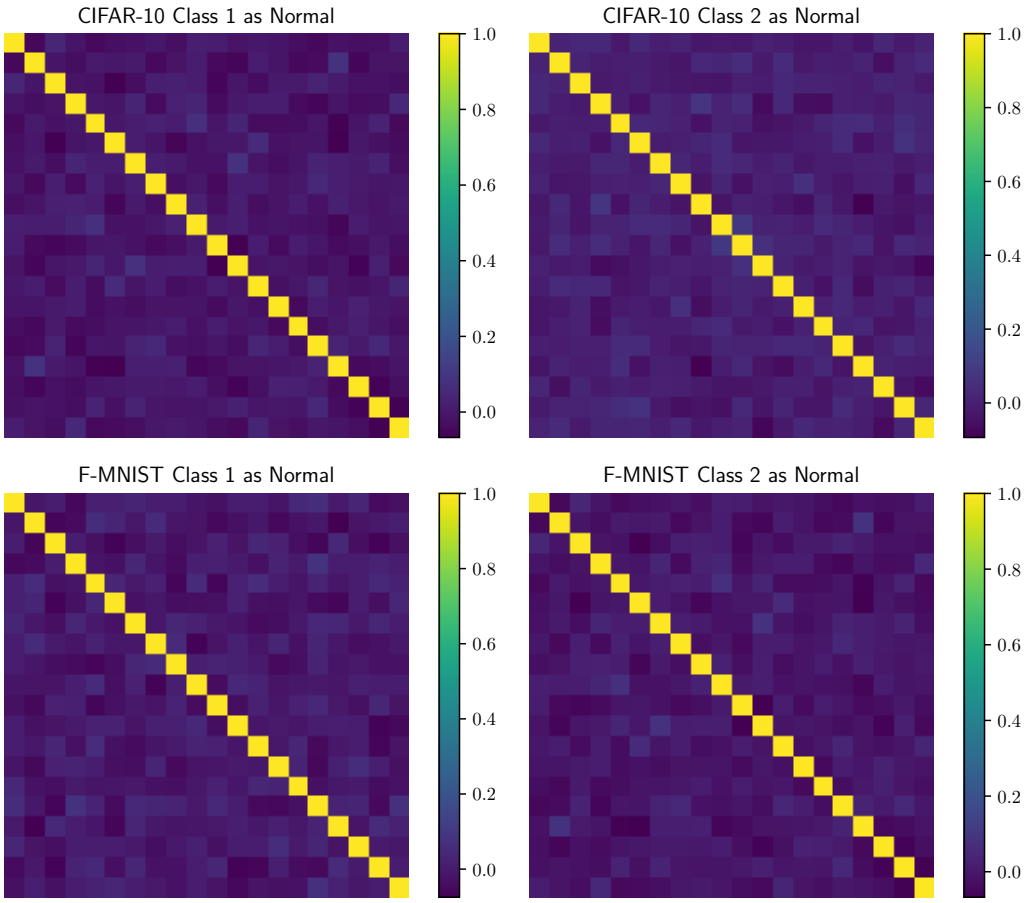

Figure 6. Anomaly score correlation matrix $\langle S(\boldsymbol{x}_i), S(\boldsymbol{x}_j)\rangle$, where $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are jointly sampled in the same query set. The result indicates that anomaly scores can be considered as approximately independent random variables.

B.2. Assumption 1

We verify Assumption 1 by showing the correlation matrix in Fig. 6, where we jointly queried 20 points with diversified querying strategy and repeated 1000 times on two classes of CIFAR-10 and F-MNIST. Then the correlation between each pair of points are computed and placed in the off-diagonal entries. For each matrix, we show the average, maximum, and minimum of the off-diagonal terms

- CIFAR-10 Class 1: -0.001, 0.103, -0.086
- CIFAR-10 Class 2: -0.001, 0.085, -0.094
- F-MNIST Class 1: -0.001, 0.081, -0.075
- F-MNIST Class 2: -0.005, 0.087, -0.067

Which shows the correlations $\langle S(\boldsymbol{x}_i), S(\boldsymbol{x}_j)\rangle$ are negligible, and the anomaly scores can be considered approximately independent random variables.

B.3. Assumption 2

We verify Assumption 2 by counting the violations, i.e., $S(\boldsymbol{x}_i) = S(\boldsymbol{x}_j)$ but $y(\boldsymbol{x}_i) \neq y(\boldsymbol{x}_j)$ (because Assumption 2 states $y_s(s_i) = y(\boldsymbol{x}_i)$ and $y_s(s_j) = y(\boldsymbol{x}_j)$, $S(\boldsymbol{x}_i) = S(\boldsymbol{x}_j)$ implies $y(\boldsymbol{x}_i) = y_s(s_i) = y_s(s_j) = y(\boldsymbol{x}_j)$. The negation is $S(\boldsymbol{x}_i) = S(\boldsymbol{x}_j)$ and $y(\boldsymbol{x}_i) \neq y(\boldsymbol{x}_j)$.). We run the experiments on both CIFAR-10 and FMNIST. We apply the "one-vs.-rest"

setup for both datasets and set the first class as normal and all the other classes as abnormal. We set the ground-truth anomaly ratio as 0.1. After the initial training, we count the pairs of data points that satisfy $S(\boldsymbol{x}_i) = S(\boldsymbol{x}_j)$ but $y(\boldsymbol{x}_i) \neq y(\boldsymbol{x}_j)$ for $i \neq j$. Our validation shows that on FMNIST, among 6666 training data points, there are 38 pairs of matching scores, and none of them have opposite labels, and on CIFAR-10, among 5555 training data points, the numbers are 21 and 3, respectively.

B.4. Contamination Ratio Estimation

Table 4. Estimated contamination ratios on CIFAR-10 and F-MNIST when $|\mathcal{Q}| = 40$ and the backbone model is NTL. The first row shows the true contamination ratio ranging from 1% to 45%. The estimations are repeated 50 times.

	1%	5%	10%	15%	20%
CIFAR-10	$0.5\% \pm 1.2\%$	$6.0\% \pm 3.3\%$	$12.0\% \pm 4.4\%$	$15.3\% \pm 4.5\%$	$18.9\% \pm 5.4\%$
F-MNIST	$1.0\% \pm 1.5\%$	$3.8\% \pm 2.3\%$	$8.7\% \pm 4.1\%$	$12.8\% \pm 5.3\%$	$19.3\% \pm 5.1\%$
	25%	30%	35%	40%	45%
CIFAR-10	$26.2\% \pm 6.0\%$	$30.6\% \pm 5.5\%$	$35.8\% \pm 6.9\%$	$42.0\% \pm 7.7\%$	$47.2\% \pm 6.7\%$
F-MNIST	$27.9\% \pm 6.4\%$	$31.8\% \pm 6.1\%$	$38.3\% \pm 6.5\%$	$43.1\% \pm 5.7\%$	$48.9\% \pm 5.6\%$

We estimate the contamination ratio by Eq. (4) under varying true ratios. This part shows the estimated contamination ratio when the query budget is $|\mathcal{Q}| = 40$. The estimations from the backbone model NTL is shown in Tab. 4. The first row contains the ground truth contamination rate, and the second and third row indicate the inferred values for two datasets, using our approach. Most estimates are withing the error bars and hence accurate. The estimation errors for low ground-truth contamination ratios are acceptable as confirmed by the sensitivity study in (Qiu et al., 2022a) which concludes that the LOE approach still works well if the anomaly ratio is mis-specified within 5 percentage points. Interestingly, we find the estimation error increases somewhat with the contamination ratio. However, a contamination ratio larger than 40% is rare in practice (most datasets should be fairly clean and would otherwise require additional preprocessing). In an anomaly detection benchmark (`https://github.com/Minqi824/ADBench`), none of the datasets have an anomaly ratio larger than 40%.

C. Baselines Details

In this section, we describe the details of the baselines in Tab. 1 in the main paper. For each baseline method, we explain their query strategies and post-query training strategies we implement in our experiment. Please also refer to our codebase for practical implementation details.

- **Rand1.** This strategy used by Ruff et al. (2019) selects queries by sampling uniformly without replacement across the training set, resulting in the queried index set $\mathcal{Q} = \{i_q \sim \text{Unif}(1, \cdots, N) | 1 \leq q \leq |\mathcal{Q}|\}$. After the querying, models are trained with a supervised loss function based on outlier exposure on the labeled data and with a one-class classification loss function on the unlabeled data,

$$L_{\text{Rand1}}(\theta) = \frac{1}{|\mathcal{Q}|} \sum_{j \in \mathcal{Q}} \big(y_j L_1^\theta(\boldsymbol{x}_j) + (1 - y_j) L_0^\theta(\boldsymbol{x}_j)\big) + \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} L_0^\theta(\boldsymbol{x}_i). \tag{6}$$

  As in SOEL both loss contributions are weighted equally. $L_{\text{Rand1}}(\theta)$ is minimized with respect to the backbone model parameters θ.

- **Rand2.** The querying strategy of Trittenbach et al. (2021) samples uniformly among the top 50% data ranked by anomaly scores without replacement. This leades to a random set of "positive" queries. After the queries are labeled, the training loss function is the same as $L_{\text{Rand1}}(\theta)$ (Eq. (6)).

- **Mar.** After training the backbone model for one epoch, this querying strategy by Görnitz et al. (2013) uses the α-quantile (s_α) of the training data anomaly scores to define a "normality region". Then the $|\mathcal{Q}|$ samples closest to the margin s_α are selected to be queried. After the queries are labeled, the training loss function is the same as $L_{\text{Rand1}}(\theta)$ (Eq. (6)). Note that in practice we don't know the true anomaly ratio for the α-quantile. In all experiment, we provide this querying strategy with the true contamination ratio of the dataset. Even with the true ratio, the "Mar" strategy is still outperformed by SOEL.

- **Hybr1.** This hybrid strategy, also used by (Görnitz et al., 2013) combines the "Mar" query with neighborhood-based diversification. The neighborhood-based strategy selects samples with fewer neighbors covered by the queried set to ensure the samples' diversity in the feature space. We start by selecting the data index $\arg\min_{1\leq i\leq N} \|s_i - s_\alpha\|$ into $\mathcal{Q}$. Then the samples are selected sequentially without replacement by the criterion

$$\underset{1\leq i\leq N}{\arg\min}\, 0.5 + \frac{|\{j \in \text{NN}_k(\phi(\boldsymbol{x}_i)) : j \in \mathcal{Q}\}|}{2k} + \beta \frac{\|s_i - s_\alpha\| - \min_i \|s_i - s_\alpha\|}{\max_i \|s_i - s_\alpha\| - \min_i \|s_i - s_\alpha\|}$$

  where the inter-sample distance is measured in the feature space and the number of nearest neighbors is $k = \lceil N/|\mathcal{Q}| \rceil$. We set $\beta = 1$ for equal contribution of both terms. After the queries are labeled, the training loss function is the same as $L_{\text{Rand1}}(\theta)$ (Eq. (6)).

- **Pos1.** This querying strategy by Pimentel et al. (2020) always selects the top-ranked samples ordered by their anomaly scores, $\arg\max_{1\leq i\leq N} s_i$. After the queries are labeled, the training loss only involves the labeled data

$$L_{\text{Pos1}}(\theta) = \frac{1}{|\mathcal{Q}|} \sum_{j\in\mathcal{Q}} \big(y_j L_1^\theta(\boldsymbol{x}_j) + (1-y_j) L_0^\theta(\boldsymbol{x}_j)\big).$$

  Pimentel et al. (2020) use the logistic loss but we use the supervised outlier exposure loss. The supervised outlier exposure loss is shown to be better than the logistic loss in learning anomaly detection models (Ruff et al., 2019; Hendrycks et al., 2018).

- **Pos2.** This approach of (Barnabé-Lortie et al., 2015) uses the same querying strategy as Pos1, but the training is different. Pos2 also uses the unlabeled data during training. After the queries are labeled, the training loss function is the same as $L_{\text{Rand1}}(\theta)$ (Eq. (6)).

- **Hybr2.** This hybrid strategy by Das et al. (2019) makes positive diverse queries. It combines querying according to anomaly scores with distance-based diversification. Hybr2 selects the initial query $\arg\max_{1\leq i\leq N} s_i$ into $\mathcal{Q}$. Then the samples are selected sequentially without replacement by the criterion

$$\underset{1\leq i\leq N}{\arg\max} \frac{s_i - \min_i s_i}{\max_i s_i - \min_i s_i} + \beta \min_{j\in\mathcal{Q}} \frac{d(\boldsymbol{x}_i, \boldsymbol{x}_j) - \min_{a\neq b} d(\boldsymbol{x}_a, \boldsymbol{x}_b)}{\max_{a\neq b} d(\boldsymbol{x}_a, \boldsymbol{x}_b) - \min_{a\neq b} d(\boldsymbol{x}_a, \boldsymbol{x}_b)}$$

  where $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = ||\phi(\boldsymbol{x}_i) - \phi(\boldsymbol{x}_j)||_2$. We set $\beta = 1$ for equal contribution of both terms. After the queries are labeled, Das et al. (2019) use the labeled set to learn a set of weights for the components of an ensemble of detectors. For a fair comparison of active learning strategies, we use the labeled set to update an individual anomaly detector with parameters θ by optimizing the loss

$$L_{\text{Hybr2}}(\theta) = \frac{1}{|\mathcal{Q}|} \sum_{j\in\mathcal{Q}} \big(y_j L_1^\theta(\boldsymbol{x}_j) + (1-y_j) L_0^\theta(\boldsymbol{x}_j)\big).$$

- **Hybr3.** This baseline by (Ning et al., 2022) uses the same query strategy as Hybr2, but differs in the training loss function,

$$L_{\text{Hybr3}}(\theta) = \frac{1}{|\mathcal{Q}| + |\mathcal{U}|} \sum_{j\in\mathcal{Q}} w_j (1-y_j) L_0^\theta(\boldsymbol{x}_j) + \frac{1}{|\mathcal{Q}| + |\mathcal{U}|} \sum_{i\in\mathcal{U}} \hat{w}_i L_0^\theta(\boldsymbol{x}_i),$$

  where $w_j = 2\sigma(d_j)$ and $\hat{w}_i = 2 - 2\sigma(d_i)$ where $\sigma(\cdot)$ is the Sigmoid function and $d_i = 10 c_d\big(||\phi(\boldsymbol{x}_i) - \boldsymbol{c}_0||_2 - ||\phi(\boldsymbol{x}_i) - \boldsymbol{c}_1||_2\big)$ where $\boldsymbol{c}_0$ is the center of the queried normal samples and $\boldsymbol{c}_1$ is the center of the queried abnormal samples in the feature space, and c_d is the min-max normalization factor.

  We make three observations for the loss function. First, $L_{\text{Hybr3}}(\theta)$ filters out all labeled anomalies in the supervised learning part and puts a large weight (but only as large as 2 at most) to the true normal data that has a high anomaly score. Second, in the unlabeled data, $L_{\text{Hybr3}}(\theta)$ puts smaller weight (less than 1) to the seemingly abnormal data. Third, overall, the weight of the labeled data is similar to the weight of the unlabeled data. This is unlike SOEL, which weighs labeled data $|\mathcal{U}|/|\mathcal{Q}|$ times higher than unlabeled data.

Algorithm 1 Training Procedure of SOEL

Input: Unlabeled training dataset $\mathcal{D}$, querying budget K
Procedure:
Train the model on $\mathcal{D}$ for one epoch as if all data were normal;
Query K data points from $\mathcal{D}$ diversely resulting in a labeled set $\mathcal{Q}$ and an unlabeled set $\mathcal{U}$;
Estimate the contamination ratio α based on $\mathcal{Q}$;
Finally train the model with $\{\mathcal{Q},\mathcal{U}\}$ until convergence:
For each iteration:
We construct a mini-batch with $\mathcal{Q}$ and a subsampled mini-batch of $\mathcal{U}$
The sample in $\mathcal{Q}$ is up-weighted with $1/|\mathcal{Q}|$ and the sample in $\mathcal{U}$ is down-weighted with weight $1/|\mathcal{U}|$
The training strategy for $\mathcal{Q}$ is supervised learning; the training strategy for $\mathcal{U}$ is LOE with the estimated anomaly ratio α.

D. Implementation Details

In this section, we present the implementation details in the experiments. They include an overall description of the experimental procedure for all datasets, model architecture, data split, and details about the optimization algorithm.

D.1. Experimental Procedure

We apply the same experimental procedure for each dataset and each compared method. The experiment starts with an unlabeled, contaminated training dataset with index set $\mathcal{U}$. We first train the anomaly detector on $\mathcal{U}$ for one epoch as if all data were normal. Then we conduct the diverse active queries at once and estimate the contamination ratio α by the importance sampling estimator Eq. (4). Lastly, we optimize the post-query training losses until convergence. The obtained anomaly detectors are evaluated on a held-out test set. The training procedure of SOEL is shown in Alg. 1.

D.2. Data Split

Image Data. For the image data including both natural (CIFAR-10 (Krizhevsky et al., 2009) and F-MNIST (Xiao et al., 2017)) and medical (MedMNIST (Yang et al., 2021)) images, we use the original training, validation (if any), and test split. When contaminating the training data of one class, we randomly sample images from other classes' training data and leave the validation and test set untouched. Specifically for DermaMNIST in MedMNIST, we only consider the classes that have more than 500 images in the training data as normal data candidates, which include benign keratosis-like lesions, melanoma, and melanocytic nevi. We view all other classes as abnormal data. Different experiment runs have different randomness.

Tabular Data. Our study includes the four multi-dimensional tabular datasets from the ODDS repository[3] which have an outlier ratio of at least 30%. . To form the training and test set for tabular data, we first split the data into normal and abnormal categories. We randomly sub-sample half the normal data as the training data and treat the other half as the test data. To contaminate training data, we randomly sub-sample the abnormal data into the training set to reach the desired 10% contamination ratio; the remaining abnormal data goes into the test set. Different experiment runs have different randomness.

Video Data. We use UCSD Peds1[4], a benchmark dataset for video anomaly detection. UCSD Peds1 contains 70 surveillance video clips – 34 training clips and 36 testing clips. Each frame is labeled to be abnormal if it has non-pedestrian objects and labeled normal otherwise. Making the same assumption as (Pang et al., 2020), we treat each frame independent and mix the original training and testing clips together. This results in a dataset of 9955 normal frames and 4045 abnormal frames. We then randomly sub-sample 6800 frames out of the normal frames and 2914 frames out of the abnormal frames without replacement to form a contaminated training dataset with 30% anomaly ratio. A same ratio is also used in the literature (Pang et al., 2020) that uses this dataset. The remaining data after sampling is used for the testing set, whose about 30% data is anomalous. Like the other data types, different experiment runs have different randomness for the training dataset construction.

[3] http://odds.cs.stonybrook.edu/
[4] http://www.svcl.ucsd.edu/projects/anomaly/dataset.htm

D.3. Model Architecture

The experiments involve two anomaly detectors, NTL and multi-head RotNet (MHRot), and three data types.

NTL on Image Data and Video Data. For all images (either natural or medical) and video frames, we extract their features by feeding them into a ResNet152 pre-trained on ImageNet and taking the penultimate layer output for our usage. The features are kept fixed during training. We then train an NTL on those features. We apply the same number of transformations, network components, and anomaly loss function $L_1^\theta(\boldsymbol{x})$, as when Qiu et al. (2022a) apply NTL on the image data.

NTL on Tabular Data. We directly use the tabular data as the input of NTL. We apply the same number of transformations, network components, and anomaly loss function $L_1^\theta(\boldsymbol{x})$, as when Qiu et al. (2022a) apply NTL on the tabular data.

MHRot on Image Data. We use the raw images as input for MHRot. We set the same transformations, MHRot architecture, and anomaly loss function as when Qiu et al. (2022a) apply MHRot on the image data.

DSVDD on Image Data. For all images (either natural or medical), we build DSVDD on the features from the penultimate layer of a ResNet152 pre-trained on ImageNet. The features are kept fixed during training. The neural network of DSVDD is a three-layer MLP with intermediate batch normalization layers and ReLU activation. The hidden sizes are $[1024, 512, 128]$.

D.4. Optimization Algorithm

Model	Dataset	Learning Rate	Epoch	Minibatch Size	τ
NTL	CIFAR-10	1e-4	30	512	1e-2
	F-MNIST	1e-4	30	512	1e-2
	MedMNIST	1e-4	30	512	1e-2
	ODDS	1e-3	100	$\lceil N/5 \rceil$	1e-2
	UCSD Peds1	1e-4	3*	192	1e-2
MHRot	CIFAR-10	1e-3	15	10	N/A
	F-MNIST	1e-4	15**	10	N/A
	MedMNIST	1e-4	15	10	N/A
Deep SVDD	CIFAR-10	1e-4	30	512	1e-2
	F-MNIST	1e-4	30	512	1e-2
	MedMNIST	1e-4	30	512	1e-2

*Hybr2, Hybr3, Pos1, and Pos2 train 30 epochs. All other methods train 3 epochs.
**SOEL train 3 epochs.

Table 5. A summary of optimization parameters for all methods.

In the experiments, we use Adam (Kingma and Ba, 2014) to optimize the objective function to find the local optimal anomaly scorer parameters θ. For Adam, we set $\beta_1 = 0.9, \beta_2 = 0.999$ and no weight decay for all experiments.

To set the learning rate, training epochs, minibatch size for MedMNIST, we find the best performing hyperparameters by evaluating the method on the validation dataset. We use the same hyperparameters on other image data. For video data and tabular data, the optimization hyperparameters are set as recommended by Qiu et al. (2022a). In order to choose τ (in Eq. (2)), we constructed a validation dataset of CIFAR-10 to select the parameter τ among {1, 1e-1, 1e-2, 1e-3} and applied the validated τ (1e-2) on all the other datasets in our experiments. Specifically, we split the original CIFAR-10 training data into a training set and a validation set. After validation, we train the model on the original training set again. We summarize all optimization hyperparameters in Tab. 5.

When training models with SOEL, we resort to the block coordinate descent scheme that update the model parameters θ and the pseudo labels $\tilde{\boldsymbol{y}}$ of unlabeled data in turn. In particular, we take the following two update steps iteratively:

- update θ by optimizing Eq. (3) given $\tilde{\boldsymbol{y}}$ fixed;
- update $\tilde{\boldsymbol{y}}$ by sovling the constrained optimization in Sec. 2.5 given θ fixed;

Upon updating $\tilde{\boldsymbol{y}}$, we use the LOE$_S$ variant (Qiu et al., 2022a) for the unlabeled data. We set the pseudo labels $\tilde{\boldsymbol{y}}$ by performing the optimization below

$$\min_{\tilde{\boldsymbol{y}}\in\{0,0.5\}^{|\mathcal{U}|}} \frac{1}{|\mathcal{U}|}\sum_{i\in\mathcal{U}} \tilde{y}_i L_1^\theta(\boldsymbol{x}_i) + (1-\tilde{y}_i)L_0^\theta(\boldsymbol{x}_i) \qquad \text{s.t.} \qquad \sum_{i=1}^{|\mathcal{U}|}\tilde{y}_i = \frac{\tilde{\alpha}|\mathcal{U}|}{2},$$

where $\tilde{\alpha}$ is the updated contamination ratio of $\mathcal{U}$ after the querying round, $\tilde{\alpha} = \big(\alpha N - \sum_{j\in\mathcal{Q}} y(\boldsymbol{x}_j)\big)/|\mathcal{U}|$, and α is computed by Eq. (4) given $\mathcal{Q}$. The solution is to rank the data by $L_0^\theta(\boldsymbol{x}) - L_1^\theta(\boldsymbol{x})$ and label the top $\tilde{\alpha}$ data abnormal (equivalently setting $\tilde{y} = 0.5$) and all the other data normal (equivalently $\tilde{y} = 0$).

When we compute the Euclidean distance in the feature space, we construct the feature vector of a sample by concatenating all its encoder representations of different transformations. For example, if the encoder representation has 500 dimensions and the model has 10 transformations, then the final feature representation has $10 \times 500 = 5000$ dimensions.

D.5. Time Complexity

Regarding the time complexity, the optimization uses stochastic gradient descent. The complexity of our querying strategy is $O(KN)$ where K is the number of queries and N is the size of the training data. This complexity can be further reduced to $O(K \log N)$ with a scalable extension of k-means++ (Bahmani et al., 2012).

E. Additional Experiments and Ablation Study

The goal of this ablation study is to show the generality of SOEL, to better understand the success of SOEL, and to disentangle the benefits of the training objective and the querying strategy. To this end, we applied SOEL to different backbone models and different data forms (raw input and embedding input), performed specialized experiments to compare the querying strategies, to demonstrate the optimality of the proposed weighting scheme in Eq. (3), and to validate the detection performance of the estimated ratio by Eq. (4). We also compared SOEL against additional baselines including semi-supervised learning frameworks and shallow anomaly detectors.

E.1. Randomness of Initialization

Random Initialization affects both the queried samples and downstream performance. To evaluate the effects, we ran all experiments 5 times with different random seeds and reported all results with error bars. In Fig. 4 we can see that the radius of the cover (a smaller radius means the queries are more diverse) does have some variance due to the random initialization. However, the corresponding results in terms of detection accuracy in Fig. 2 do have very low variance. Our interpretation is that for the CIFAR10 and F-MNIST experiments, the random initialization has little effect on detection performance.

E.2. Results with Other Backbone Models

Table 6. $|\mathcal{Q}| = 20$. AUC (%) with standard deviation for anomaly detection on six datasets (CIFAR-10, F-MNIST, Blood, OrganA, OrganC, OrganS). The backbone models are MHRot (Hendrycks et al., 2019) and Deep SVDD (Ruff et al., 2018). For all experiments, we set the contamination ratio as 10%. SOEL consistently outperforms two best-performing baselines on all six datasets.

	MHRot			Deep SVDD		
	SOEL	Hybr1	Hybr2	SOEL	Hybr1	Hybr2
CIFAR-10	**86.9±0.7**	83.9±0.1	49.1±2.0	**93.1±0.2**	89.0±0.6	91.3±1.0
F-MNIST	**92.6±0.1**	87.1±0.2	58.9±5.7	**91.4±0.5**	90.9±0.4	82.5±2.9
Blood	**83.3±0.2**	81.1±2.5	61.8±2.1	**80.2±1.1**	79.7±1.2	77.2±3.0
OrganA	**96.5±0.3**	94.1±0.3	61.1±4.8	**89.5±0.3**	87.1±0.7	71.3±3.8
OrganC	**92.1±0.2**	91.6±0.1	70.9±0.8	**87.5±0.7**	85.3±0.8	84.2±0.9
OrganS	**89.3±0.2**	88.3±0.3	68.2±0.1	**85.5±0.7**	83.4±0.3	81.2±1.3

We are interested whether SOEL works for different backbone models. To that end, we repeat part of the experiments in Tab. 2 but using an self-supervised learning model MHRot (Hendrycks et al., 2019) and a one class classification model Deep SVDD (Ruff et al., 2018) as the backbone model. We compare SOEL to two best performing baselines — Hybr1 and

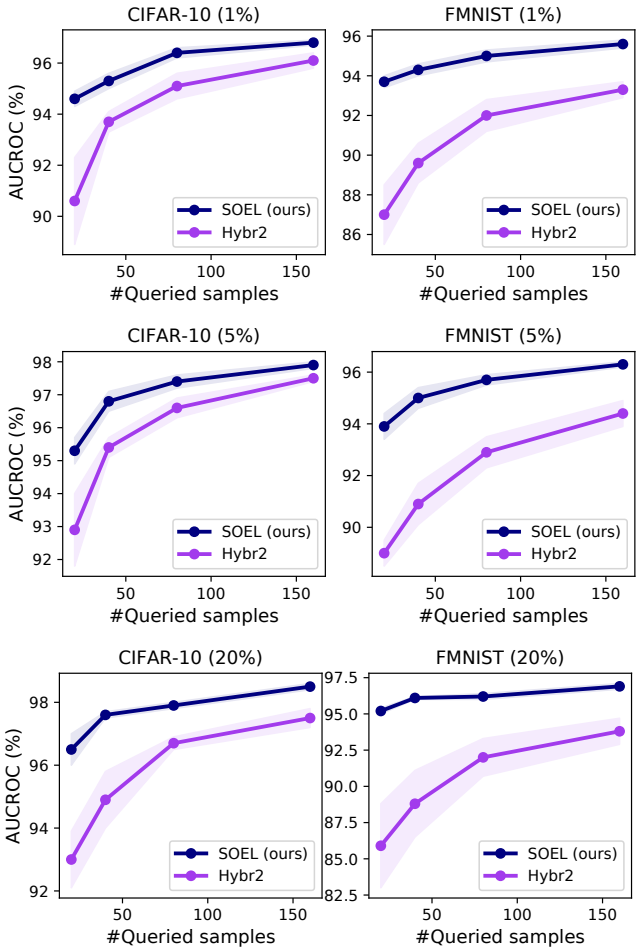

Figure 7. Running AUCs (%) with different query budgets and data contamination ratios (1%-top row, 5%-middle row, 20%-bottom row). Models are evaluated at $20, 40, 80, 160$ queries. SOEL performs the best on all three contamination ratio setups.

Hybr2. In this experiment, MHRot and Deep SVDD take different input types: while MHRot takes raw images as input, Deep SVDD uses pre-trained image features. We also set the query budget to be $|\mathcal{Q}| = 20$.

We report the results in Tab. 6. It showcases the superiority of SOEL compared to the baselines. On all datasets, SOEL significantly outperforms the two best performing baselines, Hybr1 and Hybr2, thus demonstrating the wide applicability of SOEL across anomaly detection model types.

E.3. Robustness to Anomaly Ratios

Our method works for both low anomaly ratios and high anomaly ratios. In Fig. 7, we compare SOEL against the best-performing baseline Hybr2on CIFAR-10 and FMNIST benchmarks. We vary the anomaly ratio among 1%, 5%, and 20%. On all these three anomaly ratio settings, SOEL has significantly better performance than the baseline by over 2 percentage points on average.

E.4. Disentanglement of SOEL

We disentangle the benefits of each component of SOEL and compare it to unsupervised anomaly detection with latent outlier exposure (LOE) (Qiu et al., 2022a), and to supervised active anomaly detection with k-means++ querying strategy. Both active approaches (k-means++ and SOEL) are evaluated with $|\mathcal{Q}| = 20$ labeled samples. The unsupervised approach LOE requires an hyperparameter of the assumed data contamination ratio, which we set to the ground truth value 10%.

Table 7. $|\mathcal{Q}| = 20$. AUC (%) with standard deviation for anomaly detection on CIFAR-10 and F-MNIST. For all experiments, we set the contamination ratio as 10%. SOEL mitigates the performance drop when NTL and MHRot trained on the contaminated datasets. Results of the unsupervised method LOE are borrowed from Qiu et al. (2022a).

	NTL			MHRot		
	LOE	k-means++	SOEL	LOE	k-means++	SOEL
CIFAR-10	94.9±0.1	95.6±0.3	**96.3±0.3**	86.3±0.2	64.0±0.2	**86.9±0.7**
F-MNIST	92.5±0.1	94.3±0.2	**94.8±0.4**	91.2±0.4	91.5±0.1	**92.6±0.1**

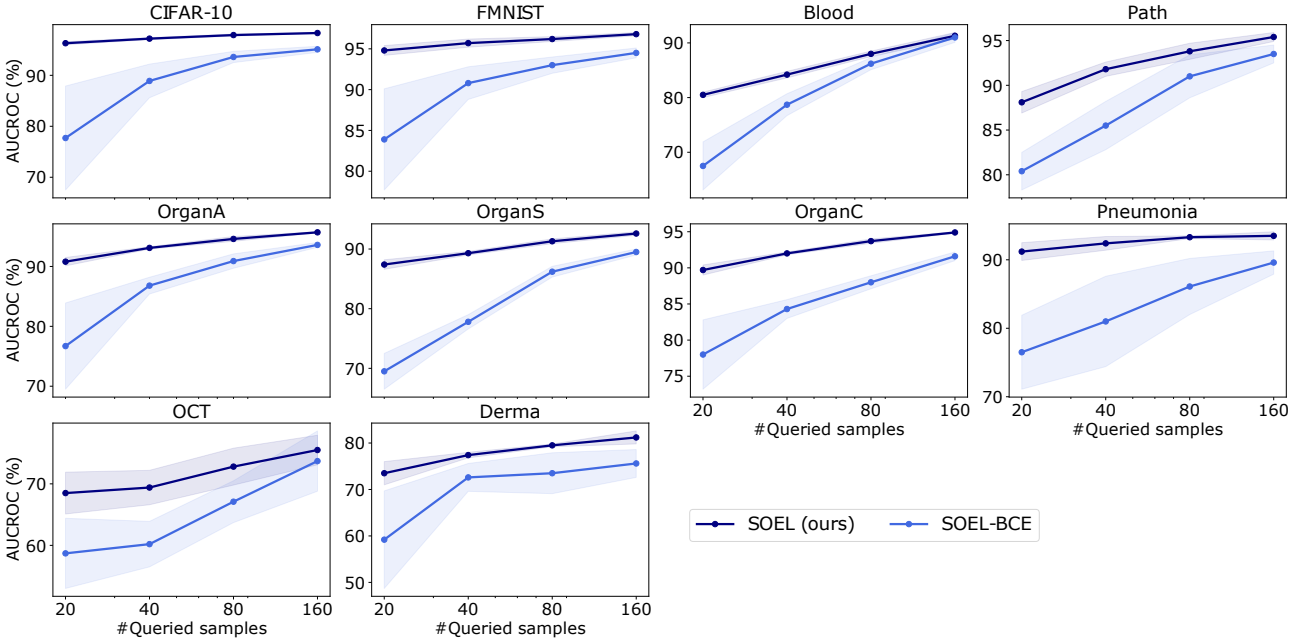

Figure 8. Running AUCs (%) with different query budgets. Models are evaluated at $20, 40, 80, 160$ queries. Deep AD model (NTL) performs significantly better than a binary classifier.

Comparing SOEL to LOE reveals the benefits of the k-means++ active approach[5]; comparing SOEL to k-means++ reveals the benefits of the unsupervised loss function in SOEL. Results in Tab. 7 show that SOEL leads to improvements for both ablation models.

E.5. Comparison to Binary Classifier

In the semi-supervised AD setup, the labeled points can be seen as an imbalanced binary classification dataset. We, therefore, perform an ablation study where we only replace deep AD backbone models with a binary classifier. All the other training and querying procedures are the same. We report the results on four different querying budget situations in Fig. 8. The figure shows that a binary classifier on all 11 image datasets falls far short of the NTL, a deep AD model. The results prove that the inductive bias (learning compact representations for normal data) used by AD models are useful for AD tasks. However, such inductive bias is lacking for binary classifiers. Especially when only querying as few as 20 points, the model can't see all anomalies. The decision boundary learned by the classifier based on the queried anomalies possibly doesn't generalize to the unseen anomalies.

E.6. Comparison to a Batch Sequential Setup

In Fig. 9, we extend our proposed method SOEL to a sequential batch active AD setup. This sequential extension is possible because our querying strategy k-means++ is also a sequential one. At each round, we query 20 points and update

[5]Notice that while LOE uses the true contamination ratio (an oracle information), SOEL only uses the estimated contamination ratio by the 20 queries.

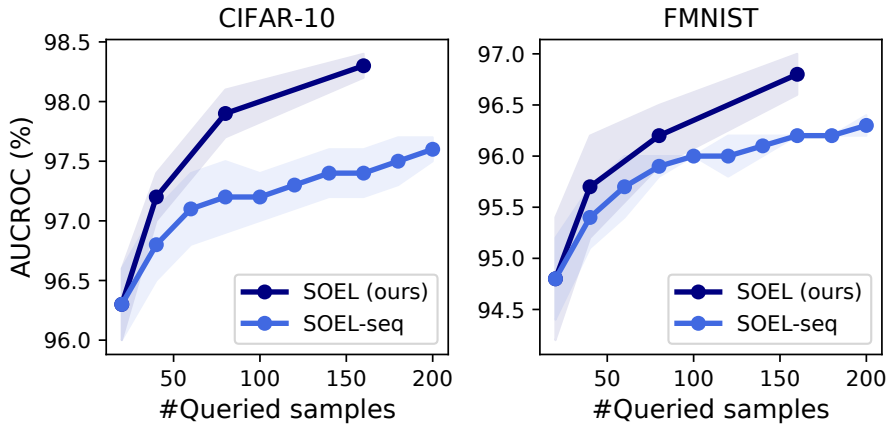

Figure 9. Running AUCs (%) with different query budgets. Models are evaluated at $20, 40, 80, 160$ queries. SOEL performs better than a sequential version.

the estimated contamination ratio. We plot this sequential version of SOEL and the original SOEL in Fig. 9 and make comparisons. The sequential version is not as effective as a single batch query of SOEL.

E.7. Comparisons of Querying Strategies

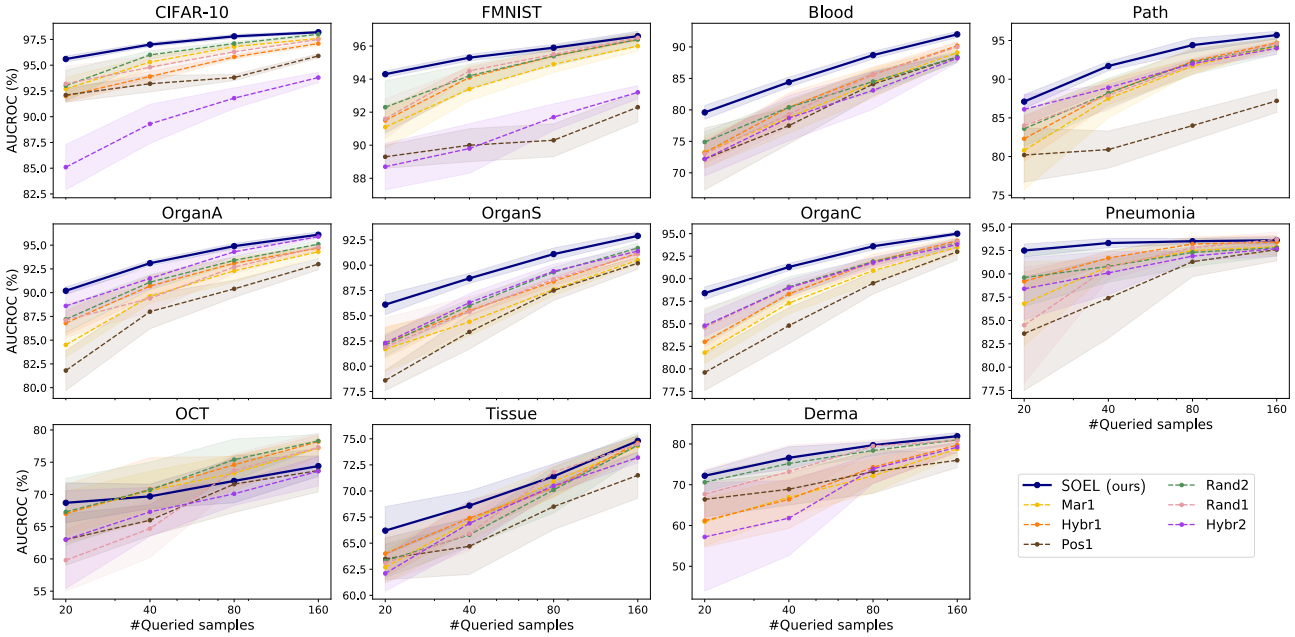

Figure 10. Ablation study on the query strategy. K-Means++ significantly outperforms other strategies for active anomaly detection on most of the datasets.

To understand the benefit of sampling diverse queries with k-means++ and to examine the generalization ability (stated in Thm. 1) of different querying strategies, we run the following experiment: We use a supervised loss on labeled samples to train various anomaly detectors. The only difference between them is the querying strategy used to select the samples. We evaluate them on all image data sets we study for varying number of queries $|\mathcal{Q}|$ between 20 and 160.

Results are in Fig. 10. On all datasets except OCT, k-means++ consistently outperforms all other querying strategies from previous work on active anomaly detection. The difference is particularly large when only few samples are queried. This also confirms that diverse querying generalizes better on the test data than other querying strategies (see additional results in Supp. A).

E.8. Ablation on Estimated Contamination Ratio

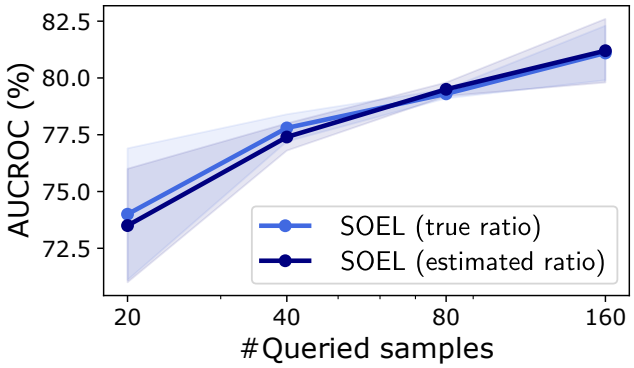

Figure 11. Model using the estimated ratio is indistinguishable from the one using the true ratio.

To see how the estimated ratio affects the detection performance, we compare SOEL to the counterpart with the true anomaly ratio. We experiment on all 11 image datasets. In Fig. 11, we report the average results for all datasets when querying $|\mathcal{Q}| = 20, 40, 80, 160$ samples. It shows that SOEL with either true ratio or estimated ratio performs similar given all query budgets. Therefore, the estimated ratio can be applied safely. This is very important in practice, since in many applications the true anomaly ratio is not known.

E.9. Ablations on Weighting Scheme

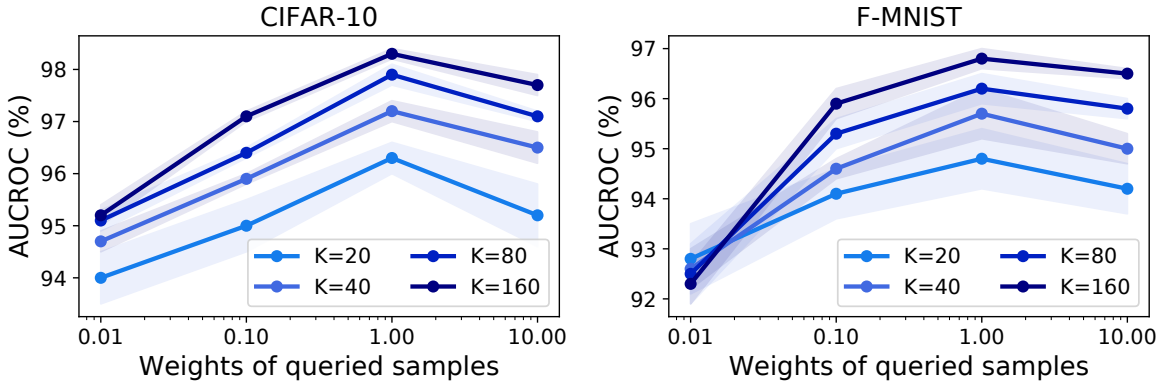

Figure 12. Ablation study on the weighting scheme in Eq. (3). With different query budgets $|\mathcal{Q}|$, the performance on image datasets degrades both upon down-weighting (0.01, 0.1) or up-weighting (10.0) the queried samples. In contrast, equal weighting yields optimal results.

We make the implicit assumption that the *averaged* losses over queried and unqueried data should be equally weighted (Eq. (3)). That means, if a fraction ϵ of the data is queried, every queried data point weights $1/\epsilon$ as much as an unqueried datum. As a consequence, neither the queried nor the unqueried data points can dominate the result.

To test whether this heuristic is indeed optimal, we added a scalar prefactor to the supervised loss in Eq. (3) (the first term) and reported the results on the CIFAR-10 and F-MNIST datasets with different query budgets (Fig. 12). A weight$<$1 corresponds to downweighting the queried term, while a weight$>$1 corresponds to upweighting it. We use the same experimental setup and backbone (NTL) as in the paper. The results are shown in Fig. 12. We see that the performance degrades both upon down-weighting (0.01, 0.1) or up-weighting (10.0) the queried samples. In contrast, equal weighting yields optimal results.

Table 8. Performance of ablation study on τ. AUROCs (%) on CIFAR-10 and F-MNIST when $|\mathcal{Q}| = 20$, the ground-truth contamination ratio is 0.1, and the backbone model is NTL.

τ	1	0.1	0.01	0.001
CIFAR-10	93.2 ± 1.7	94.5 ± 0.8	96.3 ± 0.3	95.9 ± 0.4
F-MNIST	91.8 ± 1.4	92.7 ± 1.1	94.8 ± 0.6	94.9 ± 0.2

E.10. Ablations on Temperature τ

τ (in Eq. (2)) affects the querying procedure and smaller τ makes the querying procedure more deterministic and diverse because the softmax function (in Eq. (2)) can eventually become a maximum function. We add an ablation study on different values of τ. We did experiments under the ground truth contamination ratio being 0.1 and $|Q| = 20$. As Tab. 8 shows, the smaller τ results in better AUROC results (more diverse) and smaller errors (more deterministic).

E.11. Ablations on Pseudo-label Values $\tilde{y}$

Table 9. Performance of ablation study on $\tilde{y}$. AUROC (%) on CIFAR-10 and F-MNIST when $|\mathcal{Q}| = 20$, the ground-truth contamination ratio is 0.1, and the backbone model is NTL.

$\tilde{y}$	1.0	0.875	0.75	0.625	0.5	0.25
CIFAR-10	95.3 ± 0.6	95.7 ± 0.4	95.8 ± 0.4	96.0 ± 0.5	96.3 ± 0.3	94.5 ± 0.3
F-MNIST	94.5 ± 0.5	94.5 ± 0.4	94.6 ± 0.4	94.6 ± 0.3	94.8 ± 0.6	94.0 ± 0.4

Analyzing the effects of the pseudo-label values $\tilde{y}$ is an interesting ablation study. Therefore, we perform the following experiments to illustrate the influence of different $\tilde{y}$ values. We set the ground truth contamination ratio being 0.1 and $|Q| = 20$. We vary the $\tilde{y}$ from 0.25 to 1.0 and conduct experiments. For each $\tilde{y}$ value, we run 5 experiments with different random seeds and report the AUROC results with standard deviation. It shows that $\tilde{y} = 0.5$ performs the best. While the performance of CIFAR-10 degrades slightly as $\tilde{y}$ deviates from 0.5, F-MNIST is pretty robust to $\tilde{y}$. All tested $\tilde{y}$ outperform the best baseline reported in Tab. 2.

E.12. Comparisons with Semi-supervised Learning Frameworks

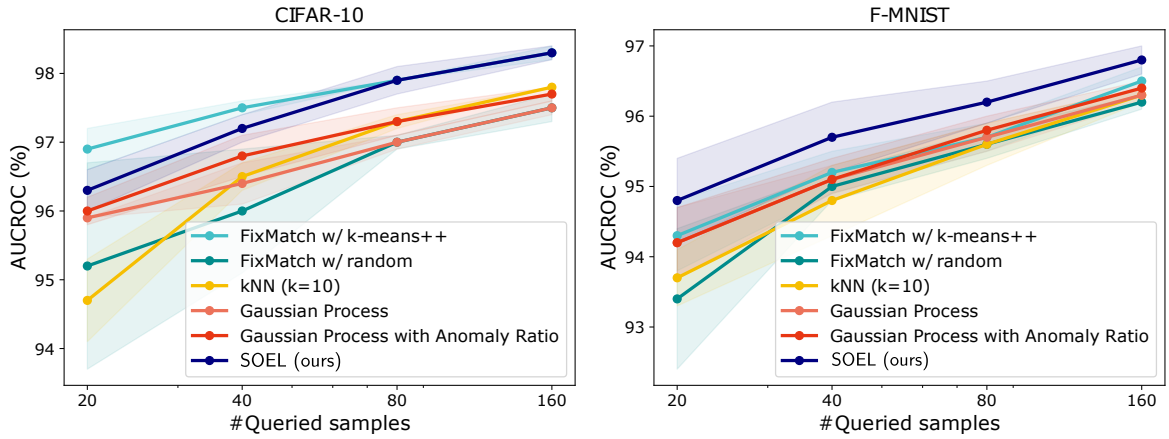

Figure 13. Comparison with semi-supervised learning fraemworks, FixMatch (Sohn et al., 2020a), k-nearest neighbors (Iscen et al., 2019), and Gaussian process (Li et al., 2018). On F-MNIST, SOEL outperforms all baselines, while on CIFAR-10, SOEL has a comparable performance with FixMatch with k-means++ querying.

SOEL exploits the unlabeled data to improve the model performance. This shares the same spirit of semi-supervised learning. We are curious about how a semi-supervised learning method performs in our active anomaly detection setup. To this end, we adapted an existing semi-supervised learning framework FixMatch (Sohn et al., 2020a) to our setup and compared with our method in Fig. 13. As follows, we will first describe the experiment results and then state the adaptation of FixMatch to

anomaly detection we made.

FixMatch, as a semi-supervised learning algorithm, regularizes the image classifier on a large amount of unlabeled data. The regularization, usually referred to consistency regularization, requires the classifier to have consistent predictions on different views of unlabeled data, thus improves the classifier's performance. FixMatch generates various data views through image augmentations followed by Cutout (DeVries and Taylor, 2017). We noticed that, although FixMatch focuses on making use of the unlabeled data, its performance is highly affected by the quality of the labeled data subset. We investigated two variants depending on how we acquire the labeled data. One is the original semi-supervised learning setting, i.e., assuming the labeled data is a random subset of the whole dataset. The other one utilizes the same diversified data querying strategy k-means++ as SOEL to acquire the labeled part. In Fig. 13, we compared the performance of the two variants with SOEL. It shows that, on natural images CIFAR10 for which FixMatch is developped, while the original FixMatch with random labeled data is still outperformed by SOEL, FixMatch with our proposed querying strategy k-means++ has a comparable performance with SOEL. However, such advantage of FixMatch diminishes for the gray image dataset F-MNIST, where both variants are beat by SOEL on all querying budgets. In addition, the FixMatch framework is restrictive and may not be applicable for tabular data and medical data, as the augmentations are specially designed for natural images.

FixMatch is designed for classification. To make it suit for anomaly detection, we adapted the original algorithm[6] and adopted the following procedure and loss function.

1. Label all training data as normal and train the anomaly detector for one epoch;
2. Actively query a subset of data with size $|\mathcal{Q}|$, resulting in $\mathcal{Q}$ and the remaining data $\mathcal{U}$;
3. Finetune the detector in a supervised manner on non-augmented $\mathcal{Q}$ for 5 epochs;
4. Train the detector with the FixMatch loss Eq. (7) on augmented $\{\mathcal{U}, \mathcal{Q}\}$ until convergence.

We denote weak augmentation of input $\boldsymbol{x}$ by $\alpha(\boldsymbol{x})$ and the strong augmentation by $\mathcal{A}(\boldsymbol{x})$. The training objective function we used is

$$\begin{aligned}\mathcal{L}_{\text{FixMatch}}(\theta) = &\frac{1}{|\mathcal{Q}|}\sum_{j\in\mathcal{Q}}\big(y_j L_1^\theta(\alpha(\boldsymbol{x}_j)) + (1-y_j)L_0^\theta(\alpha(\boldsymbol{x}_j))\big)\\ &+\frac{1}{|\mathcal{U}|}\sum_{i\in\mathcal{U}}\mathbb{1}(S(\alpha(\boldsymbol{x}_i)) < q_{0.7} \text{ or } S(\alpha(\boldsymbol{x}_i)) > q_{0.05})\big(\tilde{y}_i L_1^\theta(\mathcal{A}(\boldsymbol{x}_i)) + (1-\tilde{y}_i)L_0^\theta(\mathcal{A}(\boldsymbol{x}_i))\big)\end{aligned} \tag{7}$$

where pseudo labels $\tilde{y}_i = \mathbb{1}(S(\alpha(\boldsymbol{x}_i)) > q_{0.05}))$ and q_n is the n-quantile of the anomaly scores $\{S(\alpha(\boldsymbol{x}_i))\}_{i\in\mathcal{U}}$. In the loss function, we only use the unlabeled samples with confidently predicted pseudo labels. This is controlled by the indicator function $\mathbb{1}(S(\alpha(\boldsymbol{x}_i)) < q_{0.7}$ or $S(\alpha(\boldsymbol{x}_i)) > q_{0.05})$. We apply this loss function for mini-batches on a stochastic optimization basis.

We also extend the semi-supervised learning methods using non-parametric algorithms to our active anomaly detection framework. We applied k-nearest neighbors and Gaussian process for inferring the latent anomaly labels (Iscen et al., 2019; Li et al., 2018) because these algorithms are unbiased in the sense that if the queried sample size is large enough, the inferred latent anomaly labels approach to the true anomaly labels. For these baselines, we also queried a few labeled data with k-means++ -based diverse querying strategy and then annotate the unqueried samples with k-nearest neighbor classifier or Gaussian process classifer trained on the queried data.

Both methods become ablations of SOEL. We compare SOEL with them on CIFAR-10 and F-MNIST under various query budgets and report their results in Fig. 13. On both datasets, SOEL improves over the variant of using only queried samples for training. On F-MNIST, SOEL outperforms all ablations clearly under all query budgets, while on CIFAR-10, SOEL outperforms all ablations except for FixMatch when query budget is low. In conclusion, SOEL boosts the performance by utilizing the unlabeled samples properly, while other labeling strategies are less effective.

E.13. More Comparisons

Comparisons to kNN (Ramaswamy et al., 2000) We compared against kNN in two ways. First we confirmed that our baseline backbone model NTL is competitive with kNN, which is shown to have a strong performance on tabular

[6]We adapted the FixMatch implementation `https://github.com/kekmodel/FixMatch-pytorch`

Table 10. Comparisons with kNN method. We reported the F1-score (%) with standard error for anomaly detection on tabular datasets when the query budget $K = 10$. SOEL outperforms the kNN baseline.

	k^{th}NN	ALOE
BreastW	92.5±2.1	**93.9±0.5**
Ionosphere	88.1±1.3	**91.8±1.1**
Pima	40.5±4.7	**55.5±1.2**
Satellite	61.1±2.2	**71.1±1.7**
Average	70.6	**78.1**

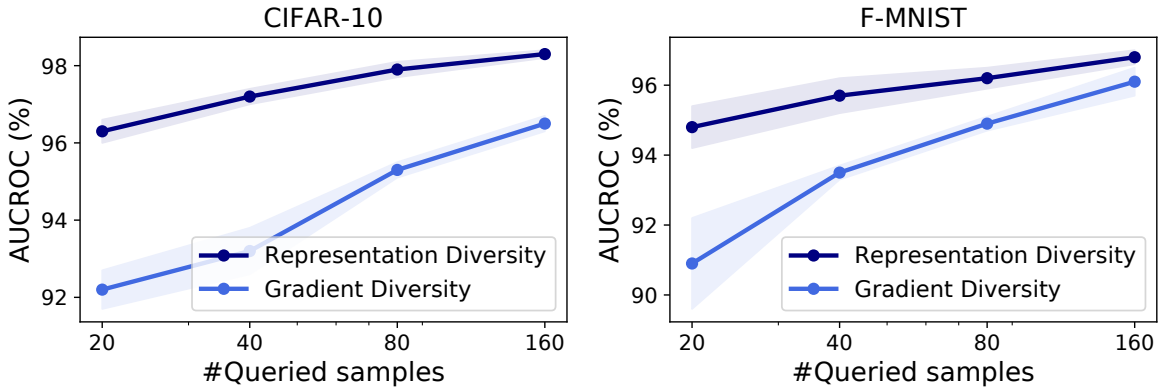

Figure 14. Comparison with gradient diversity querying strategy (BADGE) (Ash et al., 2020). The gradients wrt. the penultimate layer representation don't provide as informative queries as the representation itself, thus outperformed by our querying strategy SOEL. The true contamination ratio is 10%.

data (Shenkar and Wolf, 2022). To this end, NTL has been shown to yield 95.7% AUC on clean CIFAR-10 data, see Shenkar and Wolf, 2022, Table 1 column 1. In contrast, Qiu et al. (2022a) reported 96.2% AUC in Table 2, which is very close.

Second, we tested the performance of the kNN method on our corrupted training data set. We gave kNN the advantage of using the ground truth contamination ratio (otherwise when under-estimating this value, we saw the method degrade severely in performance).

KNN has two key hyperparameters: the number of nearest neighbors k and the assumed contamination ratio of the training set. The method uses this assumed contamination ratio when fitting to define the threshold on the decision function. In our experiments, we tried multiple values of k and reported the best testing results. Although the ground truth anomaly rate is unknown and our proposed methods don't have access to it, we gave kNN the competitive advantage of knowing the ground truth contamination ratio.

We studied the same tabular data sets as in our paper: BreastW, Ionosphere, Pima, and Satellite. We used the same procedure for constructing contaminated data outlined in our paper, where the contamination ratio was set to 10%. The results are summarized in Tab. 10.

We adopted PyOD's implementation of kNN[7] and set all the other hyperparameters to their default values ("method, radius, algorithm, leaf_size, metric, p, and metric_params"). We repeated the experiments 10 times and reported the mean and standard deviation of the F1 scores in Tab. 10. We find that our active learning framework outperforms the kNN baseline.

In more detail, the F1 scores for different values of k are listed below, where $k = 1, 2, 5, 10, 15, 20$, respectively:

- BreastW: 84.3±7.6, 86.5±3.1, 89.9±3.9, 90.7±3.4, 92.5±2.1, 91.9±1.5
- Ionosphere: 88.1±1.3, 87.6±2.6, 84.5±3.9, 75.2±2.5, 70.4±3.6, 67.4±3.4
- Pima: 34.4±3.6, 32.3±3.4, 36.9±6.4, 40.5±4.7, 35.3±3.6, 35.5±4.5

[7] https://github.com/yzhao062/pyod

- Satellite: 51.0±1.1, 53.5±0.7, 54.7±1.3, 57.4±1.8, 59.3±1.3, 61.1±2.2

Comparisons to Gradient Diversity Querying Strategy (BADGE) (Ash et al., 2020) We compared against a popular active learning method, BADGE (Ash et al., 2020), which is a diversity-driven active learning method that exploits sample-wise gradient diversity. We start with observing that BADGE doesn't work well for anomaly detection in Fig. 14, where we only replaced the objects that k-means++ works on in SOEL with gradients demanded in BADGE (Ash et al., 2020) while keeping all other settings fixed. This variant is referred to as "Gradient Diversity" while ours is denoted by "Representation Diversity". Fig. 14 shows the performance of Gradient Diversity is outperformed by a large margin, failing in querying informative samples as our Representation Diversity.

To understand which part of BADGE breaks for anomaly detection tasks, we check the gradients used by BADGE in an anomaly detection model. Before that, we start with describing how BADGE works. BADGE is developed for active learning in classification tasks. Given a pre-trained classifier, it first predicts the most likely label $\hat{y}$ (pseudo labels) for the unlabeled training data $\boldsymbol{x}$. These pseudo labels are then used to formulate a cross entropy loss $l_{CE}(\boldsymbol{x}, \hat{y})$. BADGE computes every data point's loss function's gradient to the final layer's weights as the data's representation. Upon active querying, a subset of data are selected such that their representations are diverse. In particular, the gradient to each class-specific weight W_k is $\nabla_{W_k} l_{CE}(\boldsymbol{x}, \hat{y}) = (p_k - \mathbb{1}(\hat{y} = k))\phi(\boldsymbol{x})$ where p_k is the predicted probability of being class k and $\phi(\boldsymbol{x})$ is the output of the penultimate layer. Proposition 1 of Ash et al. (2020) shows the norm of the gradient with pseudo labels is a lower bound of the one with true labels. In addition, note that the gradient is a scaling of the penultimate layer output. The scaling factor describes the predictive uncertainty and is upper bounded by 1. Therefore, the gradients are informative surrogates of the penultimate layer output of the network, as shown by the inequality

$$||\nabla_{W_k} l_{CE}(\boldsymbol{x}, \hat{y})||^2 \leq ||\nabla_{W_k} l_{CE}(\boldsymbol{x}, y)||^2 \leq ||\phi(\boldsymbol{x})||^2. \tag{8}$$

However, these properties are associated with the softmax activation function usage. In anomaly detection, models and losses are diverse and are beyond the usage of softmax activation outputs. Hence the gradients are no longer good ways to construct active queries. For example, the supervised deep SVDD (Ruff et al., 2019) uses the contrasting loss $l(\boldsymbol{x}, y) = y/(W\phi(\boldsymbol{x}) - \boldsymbol{c})^2 + (1-y)(W\phi(\boldsymbol{x}) - \boldsymbol{c})^2$ to compact the normal sample representations around center $\boldsymbol{c}$. However, the gradient $\nabla_W l(\boldsymbol{x}, y) = \big(2(1-y)(W\phi(\boldsymbol{x}) - c) - 2y(W\phi(\boldsymbol{x}) - c)^{-3}\big)\phi(\boldsymbol{x})$ is not a bounded scaling of $\phi(\boldsymbol{x})$ any more, thus not an informative surrogate of point $\boldsymbol{x}$.

E.14. NTL as a Unified Backbone Model

In Section 4 of the main paper, we have empirically compared SOEL to active-learning strategies known from various existing papers, where these strategies originally were proposed using different backbone architectures (either shallow methods or simple neural architectures, such as autoencoders). However, several recent benchmarks have revealed that these backbones are no longer competitive with modern self-supervised ones (Alvarez et al., 2022). For a fair empirical comparison of SOEL to modern baselines, we upgraded the previously proposed active-learning methods by replacing their simple respective backbones with a modern self-supervised backbone: NTL (Qiu et al., 2021)—the same backbone that is also used in SOEL.

We motivate our choice of NTL as unified backbone in our experiments as follows. Fig. 15 shows the results of ten shallow and deep anomaly detection methods (Qiu et al., 2022a; Ruff et al., 2018; Deecke et al., 2018; Golan and El-Yaniv, 2018; Hendrycks et al., 2019; Tax and Duin, 2004; Liu et al., 2008; Diederik P. Kingma, 2014; Makhzani and Frey, 2015; Sohn et al., 2020b) on the CIFAR10 one-vs.-rest anomaly detection task. NTL performs best (by a large margin) among the compared methods, including many classic backbone models known from the active anomaly detection literature (Trittenbach et al., 2021; Ruff et al., 2019; Görnitz et al., 2013; Das et al., 2019; Pimentel et al., 2020; Ning et al., 2022; Barnabé-Lortie et al., 2015).

An independent benchmark comparison of 13 methods (including nine deep methods proposed in 2018–2022) (Alvarez et al., 2022) recently identified NTL as the leading anomaly-detection method on tabular data. In their summary, the authors write: 'NeuTraLAD, the transformation-based approach, offers consistently above-average performance across all datasets. The data-augmentation strategy is particularly efficient on small-scale datasets where samples are scarce.'. Note that the latter is also the scenario where active learning is thought to be the most promising. We show the results from Alvarez et al. (2022) in Tab. 11.

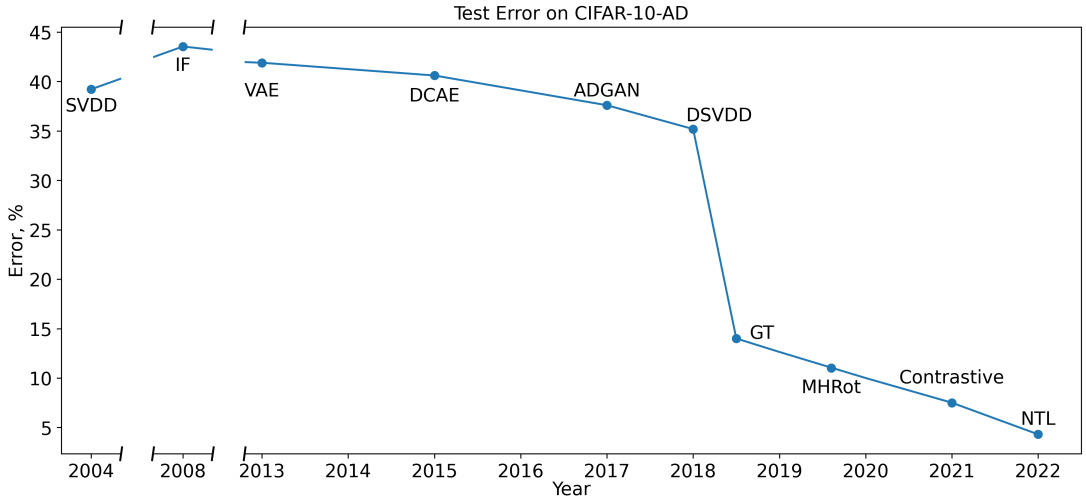

Figure 15. Error (in % of 1-AUCROC) of ten methods on CIFAR10: two shallow methods (SVDD (Tax and Duin, 2004) and IF (Liu et al., 2008)) and eight deep methods (VAE (Diederik P. Kingma, 2014), DCAE (Makhzani and Frey, 2015), ADGAN (Deecke et al., 2018), DSVDD (Ruff et al., 2018), GT (Golan and El-Yaniv, 2018), MHRot (Hendrycks et al., 2019), Contrastive (Sohn et al., 2020b), and NTL (Qiu et al., 2022a)). NTL achieves the best anomaly detection performance on CIFAR10.

Table 11. F1-scores (in %) and their standard deviations of 13 anomaly detection methods on tabular data. Results are taken from Alvarez et al. (2022). The results indicate that NTL is the state-of-the-art for tabular anomaly detection.

	KDDCUP10	NSL-KDD	IDS2018	Arrhythmia	Thyroid	Avg.
ALAD	95.9±0.7	92.1±1.5	59.0±0.0	57.4±0.4	68.6±0.5	74.6
DAE	93.2±2.0	**96.1±0.1**	**71.5±0.5**	61.5±2.5	59.0±1.5	76.3
DAGMM	95.9±1.4	85.3±7.4	55.8±5.3	50.6±4.7	48.6±8.0	67.2
DeepSVDD	89.1±2.0	89.3±2.0	20.8±11	55.5±3.0	13.1±13	53.6
DROCC	91.1±0.0	90.4±0.0	45.6±0.0	35.8±2.6	62.1±10	65.0
DSEBM-e	96.6±0.1	94.6±0.1	43.9±0.8	59.9±1.0	23.8±0.7	63.8
DSEBM-r	**98.0±0.1**	95.5±0.1	40.7±0.1	60.1±1.0	23.6±0.4	63.6
DUAD	96.5±1.0	94.5±0.2	**71.8±2.7**	60.8±0.4	14.9±5.5	67.7
MemAE	95.0±1.7	95.6±0.0	59.9±0.1	62.6±1.6	56.1±0.9	73.8
SOM-DAGMM	97.7±0.3	95.6±0.3	44.1±1.1	51.9±5.9	52.7±12	68.4
LOF	95.1±0.0	91.1±0.0	63.8±0.0	61.5±0.0	68.6±0.0	76.0
OC-SVM	96.7±0.0	93.0±0.0	45.4±0.0	**63.5±0.0**	68.1±0.0	73.3
NTL	96.4±0.2	**96.0±0.1**	59.5±8.9	60.7±3.7	**73.4±0.6**	**77.2**