

Understanding and Mitigating Data Contamination in Deep Anomaly Detection: A Kernel-based Approach

Shuang Wu¹, Jingyu Zhao², GuangJian Tian¹

¹Huawei Noah's Ark Lab

²The University of Hong Kong

wushuang.noah@huawei.com, gladsy17@connect.hku.hk, Tian.Guangjian@huawei.com

Abstract

Deep anomaly detection has become popular for its capability of handling complex data. However, training a deep detector is fragile to data contamination due to overfitting. In this work, we study the performance of the anomaly detectors under data contamination and construct a data-efficient countermeasure against data contamination. We show that training a deep anomaly detector induces an implicit kernel machine. We then derive an information-theoretic bound of performance degradation with respect to the data contamination ratio. To mitigate the degradation, we propose a contradicting training approach. Apart from learning normality on the contaminated dataset, our approach discourages learning an additional small auxiliary dataset of labeled anomalies. Our approach is much more affordable than constructing a completely clean training dataset. Experiments on public datasets show that our approach significantly improves anomaly detection in the presence of contamination and outperforms some recently proposed detectors.

1 Introduction

Anomaly detection refers to finding data points that deviate significantly from the population majority. It has a wide range of applications in network intrusion detection, disease diagnosis, and fraud identification [Chandola *et al.*, 2009]. In a typical anomaly detection setup, normal data are abundant, while anomalous data are insufficient and cannot cover all possible anomalies. The unbalanced data distribution and insufficiency of anomaly samples make standard supervised learning ineffective. Unsupervised learning thus predominates the anomaly detection models. The standard approach is to build an anomaly scoring model from the normal data. An anomaly is detectable if its anomaly score is high.

In the past two decades, a number of classic anomaly detection models have been proposed for tabular data, e.g., one-class support vector machine [Schölkopf *et al.*, 2001], local outlier factor [Breunig *et al.*, 2000], and ensemble methods such as isolation forest [Liu *et al.*, 2012]. Classic detectors determine the anomaly scores directly from fixed features.

While these detectors are efficient on tabular data with given features, the manual effort in feature engineering is not applicable for raw data from domains such as visual images, text documents, and multivariate time-series.

Deep learning techniques have recently been popular for building more powerful anomaly detectors [Pang *et al.*, 2021a]. Deep detectors (e.g., Auto-encoder [Zhou and Paffenroth, 2017], Deep SVDD [Ruff *et al.*, 2018], and GAN [Schlegl *et al.*, 2017]) wrap feature mapping and anomaly scoring into one learnable module, which automatically builds useful feature mappings that extract useful information for anomaly scoring.

The powerful modeling capacity exposes deep detectors to performance degradation due to data contamination. Recent studies, such as [Zhang *et al.*, 2017], have shown that the deep neural network can fit any training data even if they are completely random. Being oblivious of unknown anomalies mixed in the training dataset, deep anomaly detectors can be easily misguided to learn anomalies as normal ones and thus unable to detect similar anomalies. To alleviate the degradation, the training data should be free from anomaly contamination. However, manually constructing a large-scale training dataset free of anomalies can be extremely costly.

In this work, we study deep anomaly detectors in presence of training data contamination. We first show that training a deep anomaly detector induces an implicit kernel machine, which shows that the anomaly detector employs a memory-based mechanism to determine anomalies. The detector can thus be misguided by training data contaminated by anomalies. We then derive an information-theoretic degradation bound with respect to the contamination ratio. To mitigate the degradation, we propose a contradicting training approach, which utilizes an additional small dataset of labeled anomalies. Apart from the standard gradient-descent training on the contaminated dataset to memorize normal samples, we perform gradient-ascent training on the labeled anomalies to prevent the memorization of anomalous samples. Our approach only requires an additional small dataset of typical anomalies, which is much more affordable than the prohibitive costs in acquiring a large-scale normal dataset. We evaluate the effectiveness of the approach by comparing it with recently proposed semi-supervised deep anomaly detectors.

Our contributions are three-fold and summarized as follows.

- Theoretically, we develop a kernel-based framework to demonstrate the effects of training data in deep anomaly detection. Utilizing information-theoretic tools, we derive a bound characterizing the performance degradation due to training data contamination.
- Methodologically, we propose the contradicting training approach to mitigate the performance degradation. The approach is data-efficient as it only leverages a few additionally labeled anomalies to prevent the memorization of anomalies in the training dataset.
- Empirically, we conduct empirical studies on synthetic data, and publicly available datasets to validate our theoretic and methodological contributions. Our approach achieves a good detection performance and demonstrates practical effectiveness from sensitivity analysis.

2 Related Works

Anomaly detection with labeled anomalies. Motivated by the intuition that labels provide valuable prior knowledge for improving unsupervised methods, recent works have investigated using additionally labeled anomalies to improve anomaly detection. Pang et al. [Pang *et al.*, 2019] and Ruff et al. [Ruff *et al.*, 2020] incorporate labeled anomalies as additional training objectives. Zhang et al. [Zhang *et al.*, 2021] treated labeled anomalies as validation data to infer anomalies in the training set. Pang et al. [Pang *et al.*, 2021b] utilized labeled anomalies to construct an environment for reinforcement learning. While promising improvements were shown through empirical study, a recent study [Ye *et al.*, 2021] showed that whether the improvements are attainable depends on the similarity between labeled anomalies and normal data. Our work is motivated differently as we focus on the mechanism of detection performance degradation due to data contamination. A detailed comparison is in Section 5.2.

Kernel interpretation of deep learning. Our kernel-based perspective is related to the recently popular neural tangent kernel in the regime of infinitely wide neural networks [Jacot *et al.*, 2018] and the approximate kernel machines [Domingos, 2020] in a continuous-time regime. These works show an equivalence between kernel machine and gradient-based training under various setups. Our work positions the kernel equivalence in the context of anomaly detection, which illustrates the effect of data contamination in a theoretically principled framework.

Learning from noisy data. Our work is also related to classification with label noise [Hendrycks *et al.*, 2018; Diakonikolas *et al.*, 2019; Zhang *et al.*, 2020] as we treat the unlabeled large-scale dataset as noisy normal data. However, their setups consider supervised learning and assume relatively balanced label classes, which is different from anomaly detection problems. Some works [Han *et al.*, 2018; Shen and Sanghavi, 2019; Li *et al.*, 2020] assume that noisy labels induce hard-to-learn examples and attempt to lift classification accuracy by specifically regulating the model for hard-to-learn data. However, this is not applicable for anomaly detection due to under-fitting because it is difficult to distinguish hard-to-learn data from actual contamination.

3 Problem Setup

A deep anomaly detector mainly consists of an anomaly scoring model $L(x; \theta)$, where x stands for an input data and θ stands for learnable model parameters. The detector operates in two stages: training and evaluating. The training stage finds parameters that minimize anomaly scores on a training dataset $\{x_i\}_{i=1}^n$ consisting of normal samples as

$$\min_{\theta} \sum_{i=1}^n L(x_i; \theta). \quad (1)$$

In the evaluating stage, θ is obtained from solving eqn. (1). An input x is anomalous if $L(x; \theta)$ is large, and normal otherwise. The setup is general and covers a number of standard deep anomaly detection approaches such as the normality-learning framework, the end-to-end anomaly scoring framework, and a binary classification model¹.

4 Implicit Kernel in Deep Anomaly Detection

We show that training a deep anomaly detector induces an implicit kernel model. As kernels employ memory-based learning that predicts unknowns based on similarities between the input and the training data, deep detectors become ineffective in presence of training data contamination. We derive an information-theoretic bound which quantifies the performance degradation with respect to the contamination ratio.

4.1 A Representer Theorem

Training a deep anomaly detector requires solving eqn. (1). This is typically done by using the gradient descent algorithm or its variants to update the parameters iteratively as

$$\theta_{k+1} = \theta_k - \eta \sum_{i=1}^n \nabla_{\theta} L(x_i; \theta_k), \quad (2)$$

where η is the learning rate, which is typically small, and x_i is the i -th training data point. The gradient-based training implies a kernel-based perspective for predicting anomalies.

Theorem 1 (Kernel-based detector). *Assume that θ is obtained via the **gradient-based training** in eqn. (2). For sufficiently small η , the anomaly score of an input can be approximated by*

$$L(x; \theta_T) \approx L(x; \theta_0) + \sum_{i=1}^n K(x, x_i),$$

$$\text{where } K(x, x_i) = - \int_0^T \langle \nabla_{\theta} L(x; \theta_t), \nabla_{\theta} L(x_i; \theta_t) \rangle dt.$$

The theorem implies that deep detectors employ memory-based learning and similar results on supervised learning was reported in [Domingos, 2020]. In particular, a deep detector detects anomalies by using $K(\cdot, \cdot)$ to compare the input with the whole training dataset. The kernel-sum $\sum_{i=1}^n K(x, x_i)$ reflects the correlations between the input and training data. If an input is anomalous, it is not correlated with the training data and thus yields a large L with $\sum_{i=1}^n K(\cdot, x_i)$ being

¹One can set L as the final logit output.

small, which in turn becomes detectable. Otherwise, a normal input similar to the training data yields a small L with $\sum_{i=1}^n K(\cdot, x_i)$ being large, which is separable from the large L due to anomalies. This mechanism allows a deep detector to detect inputs that share no similarities with the training dataset. However, if anomalies contaminate the training dataset, the detector cannot detect inputs that are similar to those contaminations.

4.2 Information-theoretic Bound

Data contamination is widely observed in large-scale datasets. There can be up to 3% incorrect labels in human-validated datasets [Northcutt *et al.*, 2021], let alone data that are automatically collected from real-world production scenarios. Although relatively moderate label noise is tolerable for classification tasks [Rolnick *et al.*, 2017], we will show that a small contamination ratio in the training dataset significantly degrades anomaly detectability.

A maximum likelihood model. Most machine learning problems can be approached in a probabilistic paradigm and the maximum likelihood estimation is one popular technique from this perspective. We consider deep anomaly detection in this framework. In particular, training a deep neural network to minimize anomaly scores on a data sampled from the distribution P_{train} is equivalent to finding a model π that maximizes the likelihood of the sampled data as

$$(\text{training}) \quad \max_{\pi} \mathbb{E}_{x \sim P_{\text{train}}} [\log \pi(x)]. \quad (3)$$

The log-likelihood can be interpreted as the normality score of the data point. The average normality score over input data distribution P_{eval} is then

$$(\text{evaluating}) \quad \mathbb{E}_{x \sim P_{\text{eval}}} [\log \pi(x)]. \quad (4)$$

The maximum likelihood perspective echos back Theorem 1. If P_{eval} is similar to P_{train} , the evaluated normality score will be high. Otherwise, the normality score will be small, indicating that P_{eval} contains anomalous data points.

Criterion. Both the model and the similarities between P_{train} and P_{eval} contribute to the detectability of anomalies. To quantify the detectability, we define the following quantity.

Definition 1. Let P and Q be two probabilities measures defined on a sigma-algebra $\mathcal{F}$ of subsets of the same sample space Ω . The discriminating power of a model π between the two distributions is

$$D_{\pi}(P, Q) := \mathbb{E}_P[\log \pi] - \mathbb{E}_Q[\log \pi].$$

Let P be P_{train} and Q be distributions on anomalies. If P and Q are similar, $D_{\pi}(P, Q)$ is small, showing that anomalies in Q are hard to detect. Conversely, if Q is very different from P , $D_{\pi}(P, Q)$ is large and anomalies are easy to detect.

Assumption. Let P and Q stand for the data distribution of normal and anomalous data, respectively. We consider an α -contaminated training data distribution defined as $(1 - \alpha)P + \alpha Q$. The log-likelihood maximizer in eqn. (3) is $\pi^* = P_{\text{train}} = (1 - \alpha)P + \alpha Q$. We assume that model π

has a powerful fitting capacity and the training data is sufficient to cover $(1 - \alpha)P + \alpha Q$, which ensures that $\pi = \pi^*$ is attainable.

Assumption 1 (Attainability). We assume that $\pi^* = \arg \max_{\pi} \mathbb{E}_{P_{\text{train}}} [\log \pi]$ is attainable by the model π in eqn. (3) for any $0 \leq \alpha \leq 1$ in $P_{\text{train}} = (1 - \alpha)P + \alpha Q$.

The performance degradation bound. Under Assumption 1, the model learned from any α -contaminated datasets is $\pi_{\alpha} = (1 - \alpha)P + \alpha Q$. By Jensen's inequality, we can obtain that π_0 maximizes $\mathbb{E}_P[\log \pi_{\alpha}]$ and π_1 minimizes $-\mathbb{E}_Q[\log \pi_{\alpha}]$. Considering the Taylor expansion near $\alpha = 0$ (or 1), $\mathbb{E}_P[\log \pi_{\alpha}]$ (or $-\mathbb{E}_Q[\log \pi_{\alpha}]$) behaves like a monotonically decreasing quadratic function in α . One naturally wonders whether $D_{\pi}(P, Q) = \mathbb{E}_P[\log \pi_{\alpha}] - \mathbb{E}_Q[\log \pi_{\alpha}]$ decreases like a quadratic function. The answer is affirmative as stated in the following Theorem.

Theorem 2 (Quadratic degradation). Under Assumption 1, the discriminating power $D_{\pi}(P, Q)$ degrades due to α -contamination at least at an order $O(\alpha^2)$. In particular, there exist positive constants C_0 , C_1 , and ε such that

$$\begin{aligned} D_{\pi_0}(P, Q) - D_{\pi_{\alpha}}(P, Q) \\ \geq C_0 - C_1 \cdot D_{\text{KL}}(P \parallel Q) \cdot [\alpha - (1 + \varepsilon)]^2, \end{aligned}$$

where D_{KL} stands for the Kullback-Leibler (KL) divergence.

Theorem 2 shows that the detection performance monotonically decreases at a quadratic rate with respect to the contamination ratio α (given α is between 0 and 1). The anomaly detection performance degradation due to dataset contamination is amplified.

5 Contradicting Training

We propose a contradicting training approach to mitigate the performance degradation due to training data contamination.

5.1 Proposed Approach

Theorem 1 shows that anomalies become undetectable due to the memory-based prediction. Instead of manually cleaning the contaminated memory, we aim to inject additional kernel components to increase L on anomalous inputs by

$$L_{\text{fix}}(x; \theta_T) = L(x; \theta_T) + \lambda \sum_{i=n+1}^{n+m} K(x, x_i),$$

where the additional m data points are anomalies, and $\lambda > 0$ is a constant responsible for amplifying the contribution due to the additional data points. An input similar to these additional anomalies becomes detectable as L will be large due to the added kernel term. By Theorem 1, this is achievable using eqn. (2) to train the following objective

$$\min_{\theta} L_{\text{noisy}} - \lambda \cdot L_{\text{labeled}}, \quad (5)$$

where L_{noisy} stands for the original training objective consisting of both normal and anomalous data and L_{labeled} stands for m additionally labeled anomalies.

Algorithm 1 Contradicting training for one mini-batch

Input: model $\mathcal{M}$ with parameter θ , optimizer $\mathcal{O}$
 mini-batch $\{(x_i)\}_{i=1}^s$ of training data x_i ,
 labeled index in the mini-batch $\mathcal{I} \subseteq \{1, \dots, s\}$,
 hyperparameter λ
 1: $\{L_i\}_{i=1}^s \leftarrow \mathcal{M}.\text{forward}(\{(x_i)\}_{i=1}^s)$ {forward pass}
 2: **for** j in $\mathcal{I}$ **do**
 3: $L_j \leftarrow -\lambda * L_j$ {adjust loss}
 4: **end for**
 5: $L \leftarrow \sum_{i=1}^s L_i$
 6: $\nabla_{\theta} L \leftarrow \mathcal{M}.\text{backward}()$ {backward prop.}
 7: $\mathcal{O}.\text{setp}(\nabla_{\theta} L)$ {update model}

Remark 1 (Applicability for labeled normal data). *The augmented objective in Eqn. (5) is also applicable for labeled normal data as we can further augment $+\eta \cdot L_{\text{normal.labeled}}$. Nevertheless, using labeled normal data alone only reduces predicted L for normal inputs but cannot increase L for anomalous inputs.*

Data Efficiency. The augmented objective is practical if $m \ll n$. Compared with constructing a completely clean dataset, labeling a few additional labeled anomalies is much more affordable. This is attainable as we can amplify the contribution due to labeled anomalies by increasing λ as long as the labeled data are correlated with input anomalies. The contribution of training data to L_{fix} can be decomposed into

$$\underbrace{-\sum_{i=1}^{n'} K(x, x_i)}_{\text{unlabeled normal}} - \underbrace{\sum_{i=n'+1}^n K(x, x_i)}_{\text{unlabeled anomalies}} + \underbrace{\lambda \sum_{i=n+1}^{n+m} K(x, x_i)}_{\text{labeled anomalies}},$$

where n' stands for the number of unlabeled normal data in the training data. For an anomalous input x , as long as the third term is nonzero, there exists a large enough λ yielding a large L_{fix} . Our approach is thus data-efficient as long as the labeled anomalies are correlated with the anomalies to be detected. Nevertheless, one should be cautious that the labeled data should not be correlated with the normal data. Otherwise, normal data can also be classified into anomalies due to a large λ .

Implementation. Neural network is usually trained from mini-batches. We uniformly draw all samples from the whole training dataset with the labeled anomalies and perform a standard gradient descent algorithm for the losses computed with the augmented objective in eqn. (5). Algorithm 1 shows the pseudo codes for our approach within one mini-batch.

5.2 Comparison with Similar Approaches

Our proposed approach is implemented in a semi-supervised fashion, which is related to a number of recent approaches.

Ramp loss. Jehl *et al.* (2019) proposed the ramp loss $L_{\text{noisy}} - L_{\text{labeled}}$ aiming to encourage learning desired sentences and discourage undesired ones for sentence generation tasks in natural language processing. Ours reduces to the ramp loss when $\lambda = 1$. Our experiments will show that $\lambda = 1$ is not effective for anomaly detection.

DevNet. Pang *et al.* (2019) proposed a semi-supervised deep anomaly detector that minimizes anomaly scores of normal data from a Gaussian prior distribution and, in the meantime, produces large anomaly scores for labeled anomalies. The second part corresponds to our augmented term as it also encourages large L_{fix} for anomalous inputs. However, the DevNet stops gradient-based update when L_{labeled} is greater than a threshold value, which limits the contribution of labeled anomalies to produce a larger L_{fix} .

Deep SAD. Ruff *et al.* (2020) proposed an objective $L_{\text{unlabeled}} + L_{\text{labeled.normal}} + 1/L_{\text{labeled.anomaly}}$ to incorporate both labeled normal and anomalies for deep anomaly detection. The third term $1/L_{\text{labeled.anomaly}}$ contributes to a large L_{fix} for anomalous inputs as it is monotonic with respect to our proposed additional term $-L_{\text{labeled.anomaly}}$. Nevertheless, the gradient $\nabla \frac{1}{L_{\text{labeled.anomaly}}} = -\frac{\nabla L_{\text{labeled.anomaly}}}{L_{\text{labeled.anomaly}}^2}$ induces an attenuation effect, which produces a smaller L_{fix} for anomalous inputs than ours due to the multiplication factor $1/L_{\text{labeled.anomaly}}^2$.

Binary classification. We can formulate an alternative binary classification problem by treating unlabeled training data as class zero and labeled anomalies as class one. The training objective is the binary cross-entropy loss. To apply the λ amplification, we can directly multiply it with the binary classification training loss. However, the sigmoid function in the binary cross-entropy loss yields small gradients when its input is large, which results in an attenuated L_{fix}^2 .

Summary. The above works allow labeled anomalies in training as we do. However, they are motivated by incorporating labeled data into unsupervised training tasks, which is different from ours motivation in mitigating training data contamination. Consequently, they are not specifically designed to produce large L_{fix} , which makes them less efficient in mitigating contamination.

6 Experimental Study

We validate our theoretic findings and the effectiveness of our contradicting training approach using experiments. We perform experiments on one synthetic and four public datasets. Further details of experiment setups are in the appendix.

6.1 Datasets

Synthetic data. We synthesize two types of random monotonic sequences. One type is increasing and the other is decreasing. We construct an anomaly detection task based on the prediction error of interpolation. Each sequence consists of ten points. Six points are the input and the remainings are the output to be predicted. For this task, the increasing sequences are normal and the decreasing ones are anomalous. The interpolation is achieved with a multilayer perceptron.

Time-series. We use two public datasets the ECG 5000 electrocardiogram dataset³ and the HAR Human Activity

²The attenuation effect corresponds to the ‘‘gravitational pull’’, which is recently studied for the softmax function [Mei *et al.*, 2020].

³https://www.cs.ucr.edu/~eamonn/time_series_data_2018/

	train (normal)	test (normal/anomalies)
synthetic	5000	1000/1000
ECG	2500	100/100
HAR	5000	500/500
MNIST	10000	1000/9000

Table 1: Train/test split for all datasets. The row “synthetic” stands for our constructed synthetic dataset.

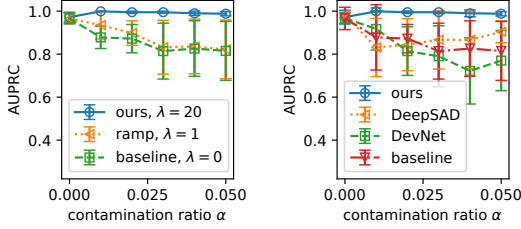


Figure 1: Results on the synthetic dataset. The markers on the lines stand for the mean and the error bars indicate standard deviations.

Recognition dataset⁴ to evaluate anomaly detection on time-series data. The ECG 5000 dataset contains five classes of the electrocardiogram. We use the normal electrocardiogram as the normal data and the R-on-T electrocardiogram as the anomalies. The HAR dataset contains six classes of human activities. We use four classes (walk, sit, stand, and lay) as normal data and the other two (upstairs and downstairs) as anomalies. The anomalies are detected according to the autoregressive prediction error computed from the output of a long short-term memory (LSTM) neural network.

Image classification. We use the MNIST dataset⁵ and the Fashion-MNIST⁶ dataset to evaluate anomaly detection on image data. Both image datasets contain ten classes and we use the standard one-v.s.-all setting to perform anomaly detection experiments. The anomalies are detected according to the reconstruction error computed from the output of a convolutional autoencoder. Model details are in the appendix.

Training/test split. Table 1 summarizes the training/test data split for each dataset. The setup of FashionMNIST is the same as that of MNIST. The size numbers for MNIST (and FashionMNIST) are the approximate numbers since the original dataset is not divided equally for each class. The number of anomaly contamination in the training set and the number of labeled anomalies for training will be specified when we present the corresponding experimental results.

6.2 Detection Metrics

Both anomaly recall rate (#recalled anomalies/#total anomalies) and the associated precision (#recalled anomalies/# recalled samples) are critical for evaluating anomaly detectors. To compare models trained with different setups and

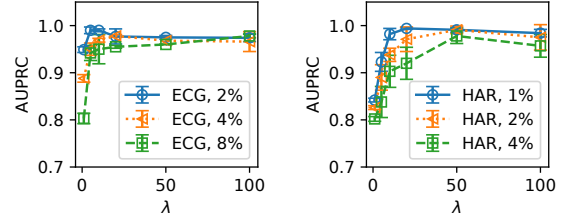
⁴<https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones>

⁵<http://yann.lecun.com/exdb/mnist/>

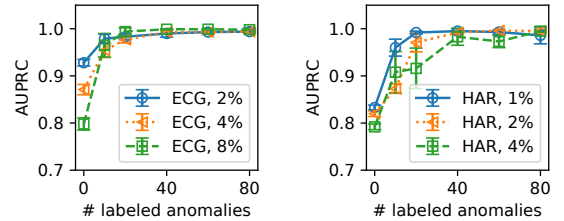
⁶<https://github.com/zalandoresearch/fashion-mnist>

Dataset	labeled	baseline	DeepSAD	DevNet	Ours
ECG	20	0.867 ±0.006	0.884 ±0.087	0.893 ±0.009	0.980 ±0.005
	50	0.860 ±0.011	0.886 ±0.062	0.925 ±0.009	0.994 ±0.002
HAR	20	0.817 ±0.003	0.330 ±0.002	0.802 ±0.013	0.969 ±0.026
	50	0.819 ±0.005	0.329 ±0.001	0.818 ±0.007	0.995 ±0.003

Table 2: Time-series results. The metric is the mean and the std. of AUPRCs. The ECG and HAR dataset have 100 and 200 additional anomalies mixed in the training dataset.



(a) Hyperparameter λ . The number of labeled anomalies is 40.



(b) Number of labeled anomalies. The hyperparameter λ is 20.

Figure 2: Sensitivity of AUPRC to the hyperparameter λ and the number of labeled anomalies. The line and the error bar indicate the mean and standard deviations from 10 random seeds, respectively. The percentiles indicate the contamination ratios for each setup.

approaches in a threshold-independent way, we evaluate the performance of an anomaly detection model based on the area under the precision-recall curve (AUPRC), which approximates the average precision weighted by the associated recall. Metrics such as **precision** rate, **recall** rate, and **F1** score (the harmonic mean of precision and recall) are not ideal since they depend on the choice of thresholds. Receiver operating characteristic (ROC) concerns mostly about samples that are determined as anomalies, which is less informative than AUPRC for anomaly detection tasks.

6.3 Algorithms in Comparison

We compare our approach with the **baseline** approach, **DeepSAD**, and **DevNet**. The details and hyperparameter settings are the same in the previous synthetic experiments. The **baseline** approach trains on contaminated training datasets without using additionally labeled anomalies. The hyperparameters of **DevNet** are selected according to the recommended value in the original papers. We add the amplification factor λ to **DeepSAD** and keep it the same as our approach.

MNIST					FashionMNIST				
class	baseline	DeepSAD	DevNet	Ours	class	baseline	DeepSAD	DevNet	Ours
0	0.982±0.000	0.824±0.000	0.983±0.001	0.992±0.005	T-shirt	0.960±0.001	0.887±0.001	0.960±0.002	0.973±0.003
1	1.000±0.000	0.985±0.006	1.000±0.000	0.996±0.001	Trouser	0.993±0.001	0.941±0.010	0.993±0.000	0.990±0.001
2	0.957±0.003	0.865±0.001	0.960±0.002	0.981±0.014	Pullover	0.943±0.002	0.872±0.037	0.945±0.002	0.986±0.001
3	0.971±0.001	0.883±0.000	0.974±0.001	0.975±0.005	Dress	0.966±0.001	0.917±0.002	0.968±0.001	0.982±0.002
4	0.979±0.001	0.928±0.000	0.979±0.001	0.995±0.001	Coat	0.957±0.002	0.837±0.000	0.959±0.002	0.977±0.001
5	0.976±0.001	0.919±0.000	0.976±0.001	0.983±0.009	Sandal	0.920±0.002	0.978±0.000	0.923±0.003	0.991±0.000
6	0.980±0.002	0.888±0.002	0.982±0.001	0.986±0.020	Shirt	0.943±0.002	0.898±0.025	0.944±0.002	0.963±0.005
7	0.986±0.000	0.937±0.000	0.986±0.001	0.996±0.001	Sneaker	0.986±0.001	0.966±0.002	0.987±0.002	0.990±0.001
8	0.955±0.003	0.870±0.000	0.959±0.003	0.942±0.036	Bag	0.923±0.002	0.879±0.019	0.924±0.001	0.988±0.002
9	0.980±0.001	0.924±0.000	0.981±0.001	0.989±0.001	AnkleBoot	0.973±0.001	0.897±0.000	0.974±0.001	0.990±0.003

Table 3: Anomaly detection performance (mean and standard deviations of AUPRC from 10 seeds) on (Fashion-)MNIST. We randomly sample 500 anomalies to contaminate all training sets and randomly sample another 100 anomalies as labeled data.

6.4 Results and Discussions

We run experiments for the five datasets and record the associated AUPRCs. For each configuration, we repeat the experiment 10 times with 10 different random seeds.

Synthetic data. We vary the contamination ratio α from 0 to 5% and compare different anomaly detectors. Figure 1 shows the AUPRCs in each setup. The curve for $\lambda = 0$ on the left panel shows that data contamination indeed causes a quadratic-like AUPRC decay. While our approach with $\lambda = 20$ reduces the performance degradation, the ramp loss [Jehl *et al.*, 2019] fails, which indicates the importance of setting a large λ . The right panel shows that our approach outperforms other recently proposed anomaly detectors.

Time-series. Table 2 shows the AUPRCs (means and standard deviations) on the time-series datasets. We contaminate ECG and HAR with 100 and 200 anomalies, respectively. For both datasets, we evaluate the detection performance with 20 labeled anomalies and 50 labeled anomalies. Our approach not only significantly improves the unsupervised baseline approach, but also outperforms other approaches. Moreover, the gradient attenuation issue is significant as the performance of DeepSAD is not very stable (better than the baseline on the synthetic dataset and the ECG 5000 dataset but much worse than the baseline in HAR).

Image data. Table 3 shows the AUPRCs (means and standard deviations) for the MNIST and Fashion-MNIST experiments. For all experiment configurations, we contaminate the training dataset with 500 randomly sampled anomalies and use another 100 randomly sampled anomalies for labeling. Our approach achieves the best performance in most of the experiments. DevNet is better than ours in a few configurations, while DeepSAD is always worse than ours. This again shows that the gradient attenuation due to the extra multiplication is harmful to anomaly detection tasks.

6.5 Sensitivity Analysis

The performance of our approach depends on the hyperparameter λ and the number of labeled anomalies. Both quantities are critical for practical applications. A desirable anomaly detector is expected to attain high AUPRCs from a large interval of λ without using too many labeled anomalies. The exact effective range of λ and the number of labeled anomalies depend on the specific dataset.

We use the two time-series datasets to investigate the practical effectiveness of our approach. We repeat each experiment using 10 random seeds. The normal training dataset is the same as before. The ECG 5000 dataset is contaminated with 2, 4, 8% contamination ratios. The ratio for the HAR dataset is 1, 2, 4%. We evaluate the AUPRCs by varying λ and the number of additionally labeled anomalies.

Figure 2 shows the mean and standard deviations (indicated in the height of the error bar) of the AUPRCs. The effective interval of λ to produce a high AUPRC is wide in general. When the number of labeled anomalies is sufficient to cover the contamination (i.e., configurations with smaller contamination ratios), a larger λ is always not worse than a smaller one. When the number of labeled anomalies is insufficient, the AUPRC first increases and then decreases while λ increases. Meanwhile, as the number of labeled anomalies inject knowledge about contamination, a large number of labeled anomalies always improve AUPRC despite that a relatively small one suffices. This validates the data efficiency.

7 Discussion and Conclusion

We developed a theoretic framework to explain the performance degradation of deep anomaly detectors due to training data contamination. We showed that the gradient-based training induced a kernel machine, which indicated a memory-based detection mechanism in deep detectors. We further showed that, given enough memory capacity, the detection performance degraded at a quadratic rate with respect to the contamination ratio. To mitigate the degradation without cleaning the training dataset, we proposed a contradicting training approach that leveraged the information of an additional small dataset of labeled anomalies.

The effectiveness of an anomaly detector relies on both the training data quality and the structure of the neural network. Our work is based on the assumption that a good neural network is available. To detect difficult unknown anomaly, techniques such as domain adaptation [Dong *et al.*, 2020] should be considered. We focus on improving anomaly detection in the presence of contamination of the dataset. Our empirical results showed that effectiveness of our approach on various public datasets. Besides the public dataset, we have also validated our approach for real-world telecommunication signaling data, which is reported in [Wu *et al.*, 2021].

References

- [Breunig *et al.*, 2000] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. LOF: identifying density-based local outliers. In *SIGMOD*, pages 93–104, 2000.
- [Chandola *et al.*, 2009] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Computing Surveys*, 41(3):1–58, 2009.
- [Diakonikolas *et al.*, 2019] Ilias Diakonikolas, Themis Gouleakis, and Christos Tzamos. Distribution-independent PAC learning of halfspaces with Massart noise. In *NeurIPS*, 2019.
- [Domingos, 2020] Pedro Domingos. Every model learned by gradient descent is approximately a kernel machine. *arXiv preprint arXiv:2012.00152*, 2020.
- [Dong *et al.*, 2020] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *CVPR*, pages 4023–4032, 2020.
- [Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, 2018.
- [Hendrycks *et al.*, 2018] Dan Hendrycks, Mantas Mazeika, Duncan Wilson, and Kevin Gimpel. Using trusted data to train deep networks on labels corrupted by severe noise. In *NeurIPS*, 2018.
- [Jacot *et al.*, 2018] Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, 2018.
- [Jehl *et al.*, 2019] Laura Jehl, Carolin Lawrence, and Stefan Riezler. Learning neural sequence-to-sequence models from weak feedback with bipolar ramp loss. *TACL*, 7:233–248, 2019.
- [Li *et al.*, 2020] Mingchen Li, Mahdi Soltanolkotabi, and Samet Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *AISTATS*, pages 4313–4324, 2020.
- [Liu *et al.*, 2012] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *TKDD*, 6(1):1–39, 2012.
- [Mei *et al.*, 2020] Jincheng Mei, Chenjun Xiao, Bo Dai, Li-hong Li, Csaba Szepesvári, and Dale Schuurmans. Escaping the gravitational pull of softmax. In *NeurIPS*, 2020.
- [Northcutt *et al.*, 2021] Curtis G Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks. *arXiv preprint arXiv:2103.14749*, 2021.
- [Pang *et al.*, 2019] Guansong Pang, Chunhua Shen, and Anton van den Hengel. Deep anomaly detection with deviation networks. In *KDD*, page 353–362, 2019.
- [Pang *et al.*, 2021a] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys*, 54(2):1–38, 2021.
- [Pang *et al.*, 2021b] Guansong Pang, Anton van den Hengel, Chunhua Shen, and Longbing Cao. Toward deep supervised anomaly detection: Reinforcement learning from partially labeled anomaly data. In *KDD*, pages 1298–1308, 2021.
- [Rolnick *et al.*, 2017] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017.
- [Ruff *et al.*, 2018] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *ICML*, pages 4393–4402, 2018.
- [Ruff *et al.*, 2020] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *ICLR*, 2020.
- [Schlegl *et al.*, 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 *Int. Conf. Info. Proc. Medical Imaging*, pages 146–157, 2017.
- [Schölkopf *et al.*, 2001] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Williamson Robert C. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13(7):1443–1471, 2001.
- [Shen and Sanghavi, 2019] Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *ICML*, pages 5739–5748, 2019.
- [Wu *et al.*, 2021] Shuang Wu, Zhen Qin, Zhen Wang, Siwei Rao, Qiang Ye, Xingyue Quan, and Guangjian Tian. Flexible o&m for telecom networks at huawei: A language model-based approach. In *CIKM*, pages 4224–4233, 2021.
- [Ye *et al.*, 2021] Ziyu Ye, Yuxin Chen, and Haitao Zheng. Understanding the effect of bias in deep anomaly detection. In *IJCAI*, 2021.
- [Zhang *et al.*, 2017] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.
- [Zhang *et al.*, 2020] Zizhao Zhang, Han Zhang, Sercan O Arik, Honglak Lee, and Tomas Pfister. Distilling effective supervision from severe label noise. In *ICCV*, pages 9294–9303, 2020.
- [Zhang *et al.*, 2021] Huayi Zhang, Lei Cao, Peter VanNostrand, Samuel Madden, and Elke A Rundensteiner. Elite: Robust deep anomaly detection with meta gradient. In *KDD*, pages 2174–2182, 2021.
- [Zhou and Paffenroth, 2017] Chong Zhou and Randy Paffenroth. Anomaly detection with robust deep autoencoders. In *KDD*, page 665–674. ACM, 2017.