Anomaly Detection by Leveraging Incomplete Anomalous Knowledge with Anomaly-Aware Bidirectional GANs

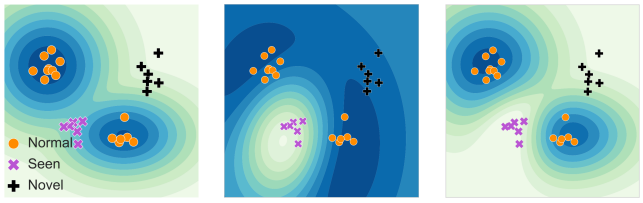
Bowen Tian[1] , **Qinliang Su**[1,2*] , **Jian Yin**[2,3] ,
[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, China
[3]School of Artificial Intelligence, Sun Yat-sen University, Guangdong, China
tianbw@mail2.sysu.edu.cn, {suqliang,issjyin}@mail.sysu.edu.cn

Abstract

The goal of anomaly detection is to identify anomalous samples from normal ones. In this paper, a small number of anomalies are assumed to be available at the training stage, but they are assumed to be collected only from several anomaly types, leaving the majority of anomaly types not represented in the collected anomaly dataset at all. To effectively leverage this kind of incomplete anomalous knowledge represented by the collected anomalies, we propose to learn a probability distribution that can not only model the normal samples, but also guarantee to assign low density values for the collected anomalies. To this end, an anomaly-aware generative adversarial network (GAN) is developed, which, in addition to modeling the normal samples as most GANs do, is able to explicitly avoid assigning probabilities for collected anomalous samples. Moreover, to facilitate the computation of anomaly detection criteria like reconstruction error, the proposed anomaly-aware GAN is designed to be bidirectional, attaching an encoder for the generator. Extensive experimental results demonstrate that our proposed method is able to effectively make use of the incomplete anomalous information, leading to significant performance gains compared to existing methods.

1 Introduction

Anomaly detection aims to identify anomalous samples from normal ones, with applications widely found in fields ranging from network security [Garcia-Teodoro *et al.*, 2009], financial fraud detection [Abdallah *et al.*, 2016], industrial damage detection to medical diagnosis [Fernando *et al.*, 2021] etc. In anomaly detection, normal data generally refers to samples preserving some kinds of regularities (typically composed of one or several types of samples), while anomaly often lacks an explicit and clear definition. Generally, any samples that deviate significantly from the normal ones are considered as anomalies. Obviously, according to this definition, the types of anomalies are numerous and sometimes even infinite. The extreme diversity of anomalies distinguish the anomaly detection problem from other tasks and also poses a significant challenge to it.



(a) Unsupervised (b) Supervised (c) Ours

Figure 1: Illustration of different approaches to leverage the collected anomalies for anomaly detection. a) Neglecting the collected anomalies; b) Finding a decision boundary separating the collected anomalies and normal samples; c) Modeling distribution of normal samples, while ensuring low densities for collected anomalies. The darker the color is, the higher the density value is.

Existing anomaly detection methods can be roughly grouped into supervised and unsupervised categories. By assuming a full accessibility to anomalies during the training, supervised methods turn the anomaly detection problem into a classification problem. Anomalous samples are generally more difficult and expensive to collect, thus unsupervised methods that only rely on the use of normal samples are used more widely in practice. Despite it is difficult to collect anomalies from every anomaly type, collecting anomalies from one or several types is often possible. For instance, by viewing images of skin diseases as anomalies, it is easy to collect some images of commonly observed skin diseases, but impossible to collect images for all known and unknown skin diseases. Under the circumstances with incompletely collected anomalies, we can still resort to the unsupervised methods by neglecting the anomalies already collected. But this approach will definitely not lead to the optimal performance for not leveraging the available anomalous knowledge, as illustrated in Fig. 1(a). Supervised methods, on the other hand, essentially seek to find a decision boundary that can separate the normal and collected anomalous samples. However, since the collected anomalies only represent a fraction of all anomaly types, when a anomaly of unseen type is presented, the classifier may not be able to classify it correctly,

*Corresponding author.

as illustrated in Fig. 1(b).

To leverage the incompletely collected anomalies, a semi-supervised anomaly detection method SSAD is proposed in [Görnitz *et al.*, 2013], which learns a compact hyper-sphere that encompasses the normal samples, while excluding the collected anomalies outside of it. In Deep SAD [Ruff *et al.*, 2019], a deep neural network is trained to encourage the learned representations of normal samples gathering towards a common center, while those of anomalies moving away from it, increasing the discrimination between normal and anomalous samples' representations. Obviously, both methods are established on a good distance metric, which, however, is often very difficult to be found, especially in the high-dimensional data space like images. Recently, [Das *et al.*, 2020] proposed to incorporate the information of anomalies identified by an expert into a basic anomaly detector through active learning framework. However, the active learning approach requires the collected anomalies to be processed one by one, losing the merits of batch processing. Moreover, to integrate with the active learning, the basic anomaly detector used in [Das *et al.*, 2020] is relatively simple, making it not suitable to be used in high-dimensional data.

In this paper, we argue that the most natural way to address this problem is to approach it from the perspective of probabilistic distribution learning. Under this view, the problem can be reformulated as learning a distribution that can not only model the normal samples well, but also ensure low density values for the collected anomalies, as illustrated in Fig. 1(c). Inspired by recent successes of deep generative models [Goodfellow *et al.*, 2014], we propose to ground this problem on them. But different from previous generative models, we developed a generative model that can not only capture the distribution of samples from the normal dataset, but also avoid to assign probabilities for samples in the abnormal dataset. Specifically, an anomaly-aware generative adversarial network (GAN) is developed, which is able to explicitly avoid generating samples that look like the training anomalies, apart from the basic capability of generating normal samples. Moreover, due to the high computational cost of directly evaluating the density value, following the unsupervised anomaly detection methods based on GANs [Zenati *et al.*, 2018b], we make the proposed anomaly-aware GAN to be bidirectional, too. With the bidirectional structure, surrogate metrics (*e.g.*, reconstruction error) can be easily calculated to assess the abnormality of a new sample. Extensive experiments were conducted to evaluate the performance of the proposed method. The results demonstrate that the proposed method is able to exploit the collected anomalies to boost the detection performance effectively, and outperforms the comparable baselines by a remarkable margin.

2 Preliminaries of GAN-Based Unsupervised Anomaly Detection

One of the main approach for unsupervised anomaly detection is to estimate the probability distribution of normal data. Generative adversarial networks (GANs) [Goodfellow *et al.*, 2014], known for their superior capability of modeling the distribution of high-dimensional data like images, have been applied to detect anomalies. In [Schlegl *et al.*, 2017], a vanilla GAN is trained to model the normal data by playing a min-max game

$$\min_G \max_D V(D,G) = \mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}\left[\log D(\mathbf{x})\right] + \mathbb{E}_{\mathbf{z}\sim p(\mathbf{z})}\left[\log\left(1 - D(G(\mathbf{z}))\right)\right], \quad (1)$$

where $G(\cdot)$ and $D(\cdot)$ represent the generator and discriminator, respectively; $p_{data}(\mathbf{x})$ and $p(\mathbf{z})$ are the distribution of normal data and standard Gaussian distribution, respectively. After training, a joint distribution $p(\mathbf{x}, \mathbf{z})$ over the data and latent code is obtained. However, due to the prohibitive computational complexity involved in the integration $p(\mathbf{x}) = \int p(\mathbf{x}, \mathbf{z}) d\mathbf{z}$, the density $p(\mathbf{x})$ cannot be used to detect anomalies directly. Instead, [Schlegl *et al.*, 2017] proposed to use gradient descent methods to find a latent code $\hat{\mathbf{z}}$ that can explain the data $\mathbf{x}$ in the latent space, and then use the reconstruction error between original data $\mathbf{x}$ and the recovered data $G(\hat{\mathbf{z}})$ to assess the abnormality, that is, $\|\mathbf{x} - G(\hat{\mathbf{z}})\|^2$. Some other surrogate criteria are also proposed, *e.g.*, leveraging the learned discriminator [Zenati *et al.*, 2018a; Sinha *et al.*, 2020]. Since the discriminator is not trained to distinguish anomalies from normal data, these criteria generally do not perform as well as the reconstruction error.

Bidirectional-GAN-Based Methods To reduce the computational cost of searching the latent code, bidirectional GANs [Dumoulin *et al.*, 2017; Li *et al.*, 2017; Donahue *et al.*, 2017] were later used for anomaly detection, in which the associated encoder can output the latent codes directly. For a bidirectional GAN, it includes two joint distributions, the encoder-induced and generator-induced joint distribution

$$p_E(\mathbf{x}, \mathbf{z}) = p_E(\mathbf{z}|\mathbf{x}) p_{data}(\mathbf{x}), \quad (2)$$

$$p_G(\mathbf{x}, \mathbf{z}) = p_G(\mathbf{x}|\mathbf{z}) p(\mathbf{z}), \quad (3)$$

where $p_E(\mathbf{z}|\mathbf{x})$ and $p_G(\mathbf{x}|\mathbf{z})$ represent the encoder $E(\cdot)$ and generator $G(\cdot)$, respectively, both of which are parameterized by neural networks. It is proved in [Dumoulin *et al.*, 2017] that $p_E(\mathbf{x}, \mathbf{z})$ and $p_G(\mathbf{x}, \mathbf{z})$ will converge to the same distribution by playing the following min-max game

$$\min_{G,E} \max_D V(D,G,E) = \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_E(\mathbf{x},\mathbf{z})}\left[\log D(\mathbf{x}, \mathbf{z})\right] + \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_G(\mathbf{x},\mathbf{z})}[\log(1 - D(\mathbf{x}, \mathbf{z}))]. \quad (4)$$

Since $p_E(\mathbf{x}, \mathbf{z})$ is defined as $p_E(\mathbf{x}, \mathbf{z}) = p_E(\mathbf{z}|\mathbf{x}) p_{data}(\mathbf{x})$ and $p_G(\mathbf{x}, \mathbf{z}) = p_E(\mathbf{x}, \mathbf{z})$ holds after convergence, it can be easily seen that the marginal distribution of $p_G(\mathbf{x}, \mathbf{z})$ *w.r.t.* $\mathbf{x}$ must be $p_{data}(\mathbf{x})$, and $p_E(\mathbf{z}|\mathbf{x})$ can be viewed as the inference network of $p_G(\mathbf{x}, \mathbf{z})$. Thus, for a sample $\mathbf{x}$, we can use the inference network $p_E(\mathbf{z}|\mathbf{x})$ to output its latent code $\hat{\mathbf{z}}$ and then use $\hat{\mathbf{z}}$ to compute the reconstruction error.

3 Anomaly Detection with Anomaly-Aware Bidirectional GANs

3.1 Problem Description

To describe the problem clearly, we define two datasets

$$\mathcal{X}^+ \triangleq \{\mathbf{x}_1^+, \mathbf{x}_2^+, \cdots, \mathbf{x}_n^+\}, \quad (5)$$

$$\mathcal{X}^- \triangleq \{\mathbf{x}_1^-, \mathbf{x}_2^-, \cdots, \mathbf{x}_m^-\}, \quad (6)$$

where $\mathcal{X}^+$ and $\mathcal{X}^-$ denote the set of normal samples and collected anomalies, respectively. According to the characteristics of anomaly detection tasks, we assume that the normal population can be sufficiently represented by the dataset $\mathcal{X}^+$, while because of the extreme diversity of anomalies, $\mathcal{X}^-$ only contain a fraction of anomaly types and cannot be used to represent the whole abnormal population. The problem interested in this paper is to judge a testing sample $\mathbf{x}$ is anomalous or not based on the two given datasets $\mathcal{X}^+$ and $\mathcal{X}^-$. Unsupervised detection methods only leverage the normal dataset $\mathcal{X}^+$, while supervised ones leverage both but assume that $\mathcal{X}^-$ can represent the entire anomaly population. Both of the methods are not suitable for the considered circumstance. Note that the dataset $\mathcal{X}^+$ is assumed to be only composed of normal samples in the analyses below, but we will show experimentally that it could be mixed with some anomalies, leading to only a slight performance drop.

3.2 Enabling Anomaly-Awareness for Bidirectional GANs under Disjoint Supports

Deep generative models have been successfully applied to detect anomalies by learning the distribution of normal samples in $\mathcal{X}^+$, but none of them make use of the collected anomalies in $\mathcal{X}^-$. In this section, to leverage the available anomalous information given by in $\mathcal{X}^-$, an **A**nomaly-**A**ware **Bi**directional **GAN** (AA-BiGAN) is developed, which can explicitly avoid to assign probabilities for samples in $\mathcal{X}^-$. To this end, we first transform the traditional bidirectional GAN described in (4) into the framework of least-square GAN (LSGAN) [Mao *et al.*, 2017], which using least-square loss to realize the updating equations. It is known that the optimization objective (4) aims to drive the output of discriminator $D(\cdot)$ towards 1 and 0 for samples from $p_E(\mathbf{x},\mathbf{z})$ and $p_G(\mathbf{x},\mathbf{z})$, respectively, while encouraging the generator $p_G(\mathbf{x}|\mathbf{z})$ and encoder $p_E(\mathbf{z}|\mathbf{x})$ to confuse the discriminator by driving it to output 0.5. Under the LSGAN framework, least-square loss is used to replace the cross-entropy loss in (4), leading to the following updating rules

$$\begin{aligned}\min_D V(D) = &\mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_E(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-1)^2\right]\\ &+\mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_G(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-0)^2\right],\end{aligned}\tag{7}$$

$$\begin{aligned}\min_{G,E} V(G,E) = &\mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_E(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-0.5)^2\right]\\ &+\mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_G(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-0.5)^2\right],\end{aligned}\tag{8}$$

where the generator G and encoder E refer to the conditional distribution $p(\mathbf{x}|\mathbf{z})$ and $p(\mathbf{z}|\mathbf{x})$, respectively; and $p_G(\mathbf{x},\mathbf{z}) = p_G(\mathbf{x}|\mathbf{z})p(\mathbf{z})$ and $p_E(\mathbf{x},\mathbf{z}) = p_E(\mathbf{z}|\mathbf{x})p_{data}(\mathbf{x})$. Following similar proofs in LSGAN, we can obtain the lemma.

Lemma 1. *If the discriminator D, generator G and encoder E are updated according to* (7) *and* (8)*, after convergence, the joint distributions $p_G(\mathbf{x},\mathbf{z})$ and $p_E(\mathbf{x},\mathbf{z})$ will be equal, that is, $p_G(\mathbf{x},\mathbf{z}) = p_E(\mathbf{x},\mathbf{z})$.*

Proof. Please refer to the Supplementary Materials. □

Although it is known that the marginal of joint distribution $p_G(\mathbf{x},\mathbf{z})$ is equal to the training data distribution $p_{data}(\mathbf{x})$ after convergence, the method still lacks the ability of explicitly avoiding to assign probabilities for samples in $\mathcal{X}^-$. To achieve this goal, we modify the updating rules in (7) and (8) by adding an anomaly-relevant term

$$\begin{aligned}\min_D V(D) = &\mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_E^+(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-1)^2\right]\\ &+\mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_E^-(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-0)^2\right]\\ &+\mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_G(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-0)^2\right],\end{aligned}\tag{9}$$

$$\begin{aligned}\min_{G,E} V(G,E) = &\mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_E^+(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-0.5)^2\right]\\ &+\mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_E^-(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-0.5)^2\right]\\ &+\mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_G(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-0.5)^2\right],\end{aligned}\tag{10}$$

where the distributions

$$p_E^+(\mathbf{x},\mathbf{z}) \triangleq p_E(\mathbf{z}|\mathbf{x})p_{\mathcal{X}}^+(\mathbf{x}),\tag{11}$$

$$p_E^-(\mathbf{x},\mathbf{z}) \triangleq p_E(\mathbf{z}|\mathbf{x})p_{\mathcal{X}}^-(\mathbf{x});\tag{12}$$

$p_{\mathcal{X}}^+(\mathbf{x})$ and $p_{\mathcal{X}}^-(\mathbf{x})$ represent the distributions of samples in $\mathcal{X}^+$ and $\mathcal{X}^-$, respectively. Although the exact expressions of $p_{\mathcal{X}}^+(\mathbf{x})$ and $p_{\mathcal{X}}^-(\mathbf{x})$ are not known, we can easily draw samples from them. The optimal discriminator is obtained

$$D^*(\mathbf{x},\mathbf{z}) = \frac{p_E^+(\mathbf{x},\mathbf{z})}{p_E^+(\mathbf{x},\mathbf{z}) + p_E^-(\mathbf{x},\mathbf{z}) + p_G(\mathbf{x},\mathbf{z})}.\tag{13}$$

By substituting the optimal discriminator $D^*(\mathbf{x},\mathbf{z})$ into (10), it can be easily verified that under the prerequisite of disjoint support $Supp(p_{\mathcal{X}}^+(\mathbf{x})) \cap Supp(p_{\mathcal{X}}^-(\mathbf{x})) = \emptyset$, the optima is achieved if and only if $p_E^+(\mathbf{x},\mathbf{z}) = p_G(\mathbf{x},\mathbf{z})$. Therefore, we have the following theorem

Theorem 2. *Suppose $Supp(p_{\mathcal{X}}^+(\mathbf{x}) \cap Supp(p_{\mathcal{X}}^-(\mathbf{x})) = \emptyset$. If the discriminator D, generator G and encoder E are updated according to* (9) *and* (10)*, after convergence, the two distributions $p_G(\mathbf{x},\mathbf{z})$ and $p_E^+(\mathbf{x},\mathbf{z})$ will be equal, that is, $p_G(\mathbf{x},\mathbf{z}) = p_E^+(\mathbf{x},\mathbf{z})$.*

Proof. Please refer to Supplementary Materials. □

From Theorem 2, it can be seen that due to $p_E^+(\mathbf{x},\mathbf{z}) \triangleq p(\mathbf{z}|\mathbf{x})p_{\mathcal{X}}^+(\mathbf{x})$, the marginal of generative distribution $p_G(\mathbf{x},\mathbf{z})$ will converge to $p_{\mathcal{X}}^+(\mathbf{x})$, which is the same as previous GANs when $p_{data}(\mathbf{x})$ is set as $p_{\mathcal{X}}^+(\mathbf{x})$. Although the new updating rules do not change the converged distribution, they bring us a different optimal discriminator as shown in (13), which is aware of the distribution of collected anomalies. To see this, let us examine the training dynamics of the generator. Suppose that a sample $(\tilde{\mathbf{x}},\tilde{\mathbf{z}})$ is generated from the generative model $p_G(\mathbf{x},\mathbf{z})$. If the sample $(\tilde{\mathbf{x}},\tilde{\mathbf{z}})$ looks similar to the collected anomalies, that is, $p_{\mathcal{X}}^-(\tilde{\mathbf{x}},\tilde{\mathbf{z}}) \gg p_{\mathcal{X}}^+(\tilde{\mathbf{x}},\tilde{\mathbf{z}})$, according to the discriminator in (13), its output value must be very small *i.e.*, close to 0. Since the target value of discriminator is 0.5 for the training of generator, the significant discrepancy between the two values will push the generator quickly moving

away from current state. Therefore, the discriminator (13) is able to prevent the generator from generating anomaly-like samples. By contrast, it can be easily shown that when setting $p_{data}(\mathbf{x})$ as $p^+_{\mathcal{X}}(\mathbf{x})$, the optimal discriminator of previous GANs obtained by (4) or (7) and (8) is

$$\widetilde{D}^*(\mathbf{x},\mathbf{z}) = \frac{p^+_E(\mathbf{x},\mathbf{z})}{p^+_E(\mathbf{x},\mathbf{z}) + p_G(\mathbf{x},\mathbf{z})}, \tag{14}$$

which does not include the anomaly-relevant term $p^-_E(\mathbf{x},\mathbf{z})$ in the discriminator. Without this term, even if a sample that looks similar to anomalies is generated, the discriminator is not guaranteed to output a small value. As a result, the discriminator $\widetilde{D}(\cdot)$ in (14) does not have the capability of explicitly preventing the generator from generating anomaly-like samples. Although theoretically the marginal of both generative distributions, no matter which discriminator is used, will converge to $p^+_{\mathcal{X}}(\mathbf{x})$ eventually, this is established on the assumptions of infinite training data and infinite representational capacities of neural networks. However, none of these conditions can be satisfied in practice, making the learned distribution only be an approximate of the ideal distribution $p^+_{\mathcal{X}}(\mathbf{x})$. Therefore, the collected anomalies are possibly assigned with some non-negligible probabilities in the learned distribution. However, if the anomaly-aware discriminator (13) is used, this kind of possibilities could be reduced significantly, thanks to the model's capability of explicitly avoiding to generate anomaly-like samples.

3.3 Relaxing the Prerequisites

The derivation of anomaly-aware bidirectional GANs in the previous section hinges on the prerequisite of disjoint supports for distributions of normal and anomalous samples $p^+_{\mathcal{X}}(\mathbf{x})$ and $p^-_{\mathcal{X}}(\mathbf{x})$, that is, $Supp(p^+_{\mathcal{X}}(\mathbf{x})) \cap Supp(p^-_{\mathcal{X}}(\mathbf{x})) = \emptyset$. However, the disjoint condition may not always hold in practice, especially when the anomalies and normal samples share lots of commonalities. In this section, we show that this prerequisite can be removed by adjusting the target values of discriminator. Moreover, we can also show that the target values do not need to be restricted to some fixed values, but could have lots of feasible configurations. To be specific, we have the following theorem.

Theorem 3. *If the discriminator D, generator G and encoder E are updated according to*

$$\begin{aligned}\min_D V(D) = {} & \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p^+_E(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-a)^2\right] \\ & + \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p^-_E(\mathbf{x},\mathbf{z})}\left[\left(D(\mathbf{x},\mathbf{z})-\frac{a+b}{2}\right)^2\right] \\ & + \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_G(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-b)^2\right], \end{aligned} \tag{15}$$

$$\begin{aligned}\min_{G,E} V(G,E) = {} & \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p^+_E(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-c)^2\right] \\ & + \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p^-_E(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-c)^2\right] \\ & + \mathbb{E}_{(\mathbf{x},\mathbf{z})\sim p_G(\mathbf{x},\mathbf{z})}\left[(D(\mathbf{x},\mathbf{z})-c)^2\right], \end{aligned} \tag{16}$$

the two distributions $p_G(\mathbf{x},\mathbf{z})$ and $p^+_E(\mathbf{x},\mathbf{z})$ after convergence will be equal, that is, $p_G(\mathbf{x},\mathbf{z}) = p^+_E(\mathbf{x},\mathbf{z})$.

Proof. Please refer to the Supplementary Materials. □

The proof of this theorem relies on the use of Jensen inequality and is more complex than that of Theorem 2. According to this theorem, we are able to obtain a bidirectional generative model that can explicitly avoid to generate anomaly-like samples even if the disjoint supports condition does not hold. For the parameters a, b and c, theoretically they could be set arbitrary. In the experiments, we follow LSGAN and set a and b as 1 and 0, respectively. For the value of c, we find that the performance is not sensitive to its value, as demonstrated by the results in Supplementary Materials. We observed that letting it far away from a, b and $\frac{a+b}{2}$ generally deliver a slightly better performance, thus we simply set it set as $\frac{3}{4}$ for all experiments.

3.4 Detection Methods

With the anomaly-aware bidirectional GAN, we can apply it to identify anomalies for new samples. Due to the difficulties of computing the marginal densities $p_G(\mathbf{x}) = \int p_G(\mathbf{x},\mathbf{z})d\mathbf{z}$, reconstruction error is used as the surrogate criteria to replace the density value, as found in many generative-model based anomaly detection methods [An and Cho, 2015; Zenati *et al.*, 2018b]. Specifically, we employ the encoder $p_E(\mathbf{z}|\mathbf{x})$ to produce a latent code $\mathbf{z}$ for the sample $\mathbf{x}$, and then use the generator to generate a sample $\hat{\mathbf{x}}$ from the code $\mathbf{z}$. The error $\|\mathbf{x}-\hat{\mathbf{x}}\|^2$ is then used to indicate the degree of abnormality of input sample $\mathbf{x}$. It is widely observed that the vanilla bidirectional GANs do not reconstruct the input well. Thus, to increase their reconstruction ability, as proposed in [Li *et al.*, 2017], one more discriminator is added to distinguish the pairs between $(\mathbf{x},\mathbf{x})$ and $(\mathbf{x},\hat{\mathbf{x}})$, with the specific expression deferred to the Supplementary Materials. In addition to the reconstruction error, some papers [Akcay *et al.*, 2018] also proposed to use the norm of latent code $\|\mathbf{z}\|$ to detect anomalies by noticing that the codes of normal samples follow a standard normal distribution $\mathcal{N}(\mathbf{0},\mathbf{I})$. Thus, if a code's norm $\|\mathbf{z}\|$ is large, it is highly suspected of being an anomaly. In practice, we find that both of the two criteria work well, but show some differences on specific datasets. In our experiments, we simply choose the one with better performance on the validation subset for a specific dataset.

4 Related Work

With the recent advancement of deep learning, neural networks have been used to help anomaly detection [Pang *et al.*, 2021a]. Among the existing methods, a typical approach is to train a deep model to reconstruct the normal data and then employ the reconstruction error to detect anomalies [Xia *et al.*, 2015]. To achieve better and robust performance, variational auto-encoders (VAE) [Kingma and Welling, 2014] are further used in [An and Cho, 2015]. To increase the distinction between the reconstruction errors of normal and anomalous samples, a memory module [Gong *et al.*, 2019] and discriminators [Perera *et al.*, 2019] are added into the autoencoders. Another type of mainstream generative models, generative adversarial networks (GAN), are also widely used for

anomaly detection [Schlegl *et al.*, 2017]. To obtain the reconstruction error more cheaply, bidirectional GANs [Dumoulin *et al.*, 2017] are proposed to use in [Zenati *et al.*, 2018b]. In addition to reconstruction-based methods, many methods established on density estimation are also studied, *e.g.*, estimating the density distribution of normal samples using energy-based models [Zhai *et al.*, 2016] or deep Gaussian mixture models [Zong *et al.*, 2018], and then using the density value to detect anomalies. There are also many other methods that resort to the one-class classifier, which assumes that normal samples can often be encompassed by a compact hypersphere [Tax and Duin, 2004] or separated by a hyperplane [Schölkopf *et al.*, 2001]. Deep SVDD [Ruff *et al.*, 2018] is further proposed, which uses neural networks to learn discriminative representations for normal and anomalous samples and then finds a hypersphere to separate them like SVDD. In recent years, it has been pointed out that it is possible to collect a handful of anomalies before the training in practice. Early methods simply turn this problem into a binary classification problem [Kingma *et al.*, 2014], without considering the incompleteness of the collected anomalies. Differently, SSAD [Görnitz *et al.*, 2013] proposed to ground the problem on unsupervised detection method SVDD, while ensuring the available anomalous samples are outside of the hypersphere. Later, deep SAD [Ruff *et al.*, 2019] proposed to learn a mapping function that encourages the representations of normal samples gathering toward a center, while those of collected anomalies moving away from it. Although both methods achieve superior performance, these methods heavily rely on the use of a good distance metric, which is often difficult to be found in high-dimensional data. Recently, [Das *et al.*, 2020] seeks to use active learning to incorporate the anomalies identified by an expert into the existing basic anomaly detector. [Pang *et al.*, 2021b] proposed a reinforcement-learning-based method to actively seek novel types of anomalies. However, the architecture of basic anomaly detectors employed in both methods is relatively simple, impeding them from being applied to high-dimensional data with complex structure like images.

5 Experiments

5.1 Experimental Setups

Training & Evaluation Following the paper [Ruff *et al.*, 2019], for each dataset, we select one category as normal, while treating the remaining nine types as anomalies. To mimic the circumstance of incomplete anomalous information, a proportion of samples from one of the nine anomalous categories have been collected. The proposed model is trained on the normal samples and collected anomalies. The testing dataset is splitted into a validation and testing dataset with a ratio of 20% and 80%. The hyperparameters are fine-tuned on the validation dataset, please refer to the Supplementary Material for more details of training. The area under the receiver operating characteristic curve (AUROC) is employed as the performance criteria. The reported experimental results are averaged over 90 experiments (*i.e.,* 10 choices of normal categories $\times$ 9 choices of anomalous categories).

Baselines Semi-supervised anomaly detection methods are used for comparison: *SSAD* [Görnitz *et al.*, 2013], *SS-DGM* [Kingma *et al.*, 2014], *Deep SAD* [Ruff *et al.*, 2019], negative data augmentation method (*NDA*) [Sinha *et al.*, 2020], and active anomaly detection (*AAD*) [Das *et al.*, 2020] for tabular dataset. In addition, a supervised binary classifier is also trained for comparison. The performance of these methods, except NDA and AAD, are quoted from Deep SAD in [Ruff *et al.*, 2019]. The NDA and AAD papers concentrate on different scenarios as ours, thus we report their performance by running and tuning their publicized code on our experiments.

5.2 Performance and Analysis

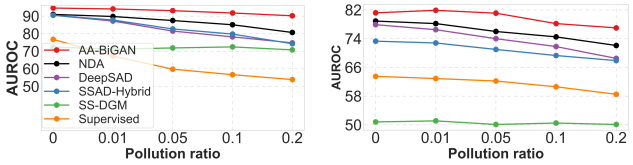
Overall Performance Table 1 shows the performance of our proposed model[1]. To investigate how the number of collected anomalies affects the performance, the performance is evaluated under different values of γ_l, where γ_l is defined as the ratio between the number of collected anomalies and the number of normal samples. When γ_l is set to 0, the semi-supervised methods are reduced to the unsupervised ones. It can be seen that as long as a very small fraction of anomalies is used, a significant performance improvement can be observed, especially on the complex dataset of CIFAR-10. And as the ratio of available anomalies further increases, a steady improvement can be observed. Under the specific case of $\gamma_l = 0.01$, the performance gains observed in our proposed method over the best baseline on MNIST, F-MNIST and CIFAR-10 are 0.6%, 2.3% and 6.8%, respectively, demonstrating that our method can exploit the partially-observed anomalies effectively to boost the overall detection performance. Moreover, it can be also seen that as the dataset of interest becomes more complex, the advantages of our proposed method become more clear. This may be attributed to the strong ability of GANs in modeling the probabilistic distribution of image data. By contrast, due to the difficulties of finding an appropriate metric to measure the distance in high-dimensional data space, distance-based detection methods, like SSAD and Deep SAD, struggle in such scenarios. As for the NDA, although it is also established on GANs, it focuses more on the generation of images, lacking the necessary bidirectional structure to produce the surrogate detection criteria like reconstruction error as in our model.

Performance with polluted normal dataset $\mathcal{X}^+$ In some scenarios, it may be difficult to obtain a clean dataset $\mathcal{X}^+$, in which a small proportion of anomalous samples may be mixed. To investigate how our proposed model performs under such circumstances, experiments are conducted under different ratios of pollution γ_p, where γ_p is defined as the ratio between the number of pollution anomalies and the number of normal samples in the training. Fig. 2 shows the detection performance of different models as a function of γ_p on different datasets, in which the γ_l is fixed as 0.05. It can be seen that the performance of almost all methods deteriorates as the pollution ratio γ_p increases from 0.0 to 0.2. However, the proposed method decreases much slower than the compared ones. This may be because the probabilistic method is more tolerant to data pollution than distance-based methods. We

[1]Code is available at https://github.com/tbw162/AA-BiGAN.

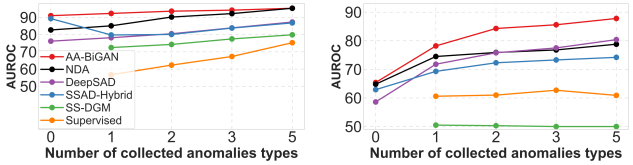
Data	γ_l	SSAD Raw	SSAD Hybrid	SS-DGM	Supervised Classifier	Deep SAD	NDA	AA-BiGAN (ours)
MNIST	.00	96.0 ± 2.9	**96.3 ± 2.5**	n.a.	n.a.	92.8 ± 4.9	86.5 ± 9.5	**96.3 ± 2.8**
	.01	96.6 ± 2.4	96.8 ± 2.3	89.9 ± 9.2	92.8 ± 5.5	96.4 ± 2.7	96.5 ± 3.7	**97.4 ± 2.1**
	.05	93.3 ± 3.6	97.4 ± 2.0	92.2 ± 5.6	94.5 ± 4.6	96.7 ± 2.4	96.8 ± 3.2	**97.8 ± 2.0**
	.10	90.7 ± 4.4	97.6 ± 1.7	91.6 ± 5.5	95.0 ± 4.7	96.9 ± 2.3	96.9 ± 3.0	**98.4 ± 1.7**
	.20	87.2 ± 5.6	97.8 ± 1.5	91.2 ± 5.6	95.6 ± 4.4	96.9 ± 2.4	97.1 ± 2.9	**98.2 ± 1.6**
F-MNIST	.00	92.8 ± 4.7	91.2 ± 4.7	n.a.	n.a.	89.2 ± 6.2	82.7 ± 11.4	**93.0 ± 4.8**
	.01	92.1 ± 5.0	89.4 ± 6.0	65.1 ± 16.3	74.4 ± 13.6	90.0 ± 6.4	90.1 ± 8.5	**94.4 ± 5.0**
	.05	88.3 ± 6.2	90.5 ± 5.9	71.4 ± 12.7	76.8 ± 13.2	90.5 ± 6.5	91.0 ± 7.1	**94.6 ± 4.2**
	.10	85.5 ± 7.1	91.0 ± 5.6	72.9 ± 12.2	79.0 ± 12.3	91.3 ± 6.0	91.4 ± 7.0	**94.7 ± 4.3**
	.20	82.0 ± 8.0	89.7 ± 6.6	74.7 ± 13.5	81.4 ± 12.0	91.0 ± 5.5	91.4 ± 7.1	**94.8 ± 4.1**
CIFAR-10	.00	62.0 ± 10.6	63.8 ± 9.0	n.a.	n.a.	60.9 ± 9.4	64.8 ± 8.2	**65.0 ± 9.4**
	.01	73.0 ± 8.0	70.5 ± 8.3	49.7 ± 1.7	55.6 ± 5.0	72.6 ± 7.4	73.4 ± 7.7	**80.2 ± 6.6**
	.05	71.5 ± 8.1	73.3 ± 8.4	50.8 ± 4.7	63.5 ± 8.0	77.9 ± 7.2	78.9 ± 7.7	**81.5 ± 6.4**
	.10	70.1 ± 8.1	74.0 ± 8.1	52.0 ± 5.5	67.7 ± 9.6	79.8 ± 7.1	80.1 ± 7.6	**83.6 ± 6.4**
	.20	67.4 ± 8.8	74.5 ± 8.0	53.2 ± 6.7	80.5 ± 5.9	81.9 ± 7.0	81.6 ± 7.3	**84.8 ± 7.2**

Table 1: Averaged AUROC under different collected-anomaly ratios γ_l.



F-MNIST CIFAR-10

Figure 2: The detection performance as a function of pollution ratio γ_p on F-MNIST and CIFAR-10 datasets.



F-MNIST CIFAR-10

Figure 3: The detection performance as a function of the number of categories of collected anomalies k_l.

can observe that the performance of NDA decreases quickly as γ_p increases. This may be because data pollution breaks its required distribution disjoint assumption in NDA.

Impact of the diversity of collected anomalies Fig. 3 shows how the performance of different methods varies as the number of categories of collected anomalies increases from 0 to 5, in which γ_l and γ_p are fixed as 0.05 and 0.1, respectively. It can be seen that the performance of all considered methods improves steadily as the number of categories increases from 0 to 5. This is consistent with our intuition since more types of anomalies are exposed to the models. However, the gain of our proposed method becomes less significant as more and more categories of anomalies are added into the training. This is easy to understand, when anomalies from five categories are used, it means anomalies from over a half of anomalous categories are accessible during the training, making the problem less challenging and hence the advantage of our method less obvious. But due to the extreme diversity of anomalies in real-world applications, the collected anomalies typically can only account for a very small fraction of all types, which suggests that the setups with small number of categories are actually more meaningful.

Data	AAD	SSAD Hybrid	Supervised Classifier	Deep SAD	AA-BiGAN (Ours)
Arrhythmia	75.8±3.2	78.3±5.1	39.2±9.5	75.9±8.7	**80.7±3.2**
Cardio	90.7±2.1	86.3±5.8	83.2±9.6	95.0±1.6	**98.0±1.2**
Satellite	77.2±4.1	86.9±2.8	87.2±2.1	**91.5±1.1**	87.4±2.3
Satimage-2	**99.9±0.1**	96.8±2.1	99.1±0.1	**99.9±0.1**	**99.9±0.1**
Shuttle	99.0±0.2	97.7±1.0	95.1±8.0	98.4±0.9	**99.1±0.1**
Thyroid	96.5±0.8	95.3±3.1	97.8±2.6	98.6±0.9	**98.9±0.1**

Table 2: AUROC on classic anomaly detection datasets.

Performance on other anomaly detection datasets We evaluated our method on six other classic anomaly detection datasets. Table 2 shows the performance of our proposed model under the scenario of $\gamma_l = 0.01$ and $\gamma_p = 0$. From the table, it can be seen our proposed model overall outperforms current baseline methods. Even on Arrhythimia dataset, which contains less than 500 samples, our model still achieve a 2% performance improvement, demonstrating the competitiveness of the proposed method on small datasets.

6 Conclusion

In this paper, we studied the problem of anomaly detection under the circumstance that a handful of anomalies are available during the training. To effectively leverage the incomplete anomalous information to help anomaly detection, an anomaly-aware GAN is developed, which is able to explicitly avoid assigning probabilities for the collected anomalies, apart from the basic capabilities of modeling the distribution of normal samples. To facilitate the computation of anomaly detection criteria like reconstruction error, the anomaly-aware GAN is designed to be bidirectional. Extensive experiments demonstrated that under the circumstance of incomplete anomalous information, our model significantly outperformed existing baseline methods.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 61806223, U1811264), Key R&D Program of Guangdong Province (No. 2018B010107005), National Natural Science Foundation of Guangdong Province (No. 2021A1515012299). This work is also sponsored by CAAI-Huawei MindSpore Open Fund.

References

[Abdallah *et al.*, 2016] Aisha Abdallah, Mohd Aizaini Maarof, and Anazida Zainal. Fraud detection system: A survey. *Journal of Network and Computer Applications*, 68:90–113, 2016.

[Akcay *et al.*, 2018] Samet Akcay, Amir Atapour-Abarghouei, and Toby P Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *ACCV*, pages 622–637, 2018.

[An and Cho, 2015] Jinwon An and Sungzoon Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015.

[Das *et al.*, 2020] Shubhomoy Das, Weng-Keen Wong, Thomas Dietterich, Alan Fern, and Andrew Emmott. Discovering anomalies by incorporating feedback from an expert. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 14(4):1–32, 2020.

[Donahue *et al.*, 2017] Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *ICLR*, 2017.

[Dumoulin *et al.*, 2017] Vincent Dumoulin, Ishmael Belghazi, Ben Poole, Alex Lamb, Martín Arjovsky, Olivier Mastropietro, and Aaron C. Courville. Adversarially learned inference. In *ICLR*, 2017.

[Fernando *et al.*, 2021] Tharindu Fernando, Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Deep learning for medical anomaly detection–a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021.

[Garcia-Teodoro *et al.*, 2009] Pedro Garcia-Teodoro, Jesus Diaz-Verdejo, Gabriel Maciá-Fernández, and Enrique Vázquez. Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security*, 28(1-2):18–28, 2009.

[Gong *et al.*, 2019] Dong Gong, Lingqiao Liu, Vuong Le, Budhaditya Saha, Moussa Reda Mansour, Svetha Venkatesh, and Anton van den Hengel. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *CVPR*, pages 1705–1714, 2019.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *NeurIPS*, 2014.

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

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *ICLR*, 2014.

[Kingma *et al.*, 2014] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *NeurIPS*, pages 3581–3589, 2014.

[Li *et al.*, 2017] Chunyuan Li, Hao Liu, Changyou Chen, Yuchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. Alice: Towards understanding adversarial learning for joint distribution matching. *NeurIPS*, pages 5495–5503, 2017.

[Mao *et al.*, 2017] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2794–2802, 2017.

[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 (CSUR)*, 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 *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1298–1308, 2021.

[Perera *et al.*, 2019] Pramuditha Perera, Ramesh Nallapati, and Bing Xiang. Ocgan: One-class novelty detection using gans with constrained latent representations. In *CVPR*, pages 2898–2906, 2019.

[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.*, 2019] Lukas Ruff, Robert A Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *ICLR*, 2019.

[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 *IPMI*, pages 146–157, 2017.

[Schölkopf *et al.*, 2001] 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.

[Sinha *et al.*, 2020] Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon. Negative data augmentation. In *ICLR*, 2020.

[Tax and Duin, 2004] David MJ Tax and Robert PW Duin. Support vector data description. *Machine learning*, 54(1):45–66, 2004.

[Xia *et al.*, 2015] Yan Xia, Xudong Cao, Fang Wen, Gang Hua, and Jian Sun. Learning discriminative reconstructions for unsupervised outlier removal. In *ICCV*, pages 1511–1519, 2015.

[Zenati *et al.*, 2018a] Houssam Zenati, Chuan Sheng Foo, Bruno Lecouat, Gaurav Manek, and Vijay Ramaseshan Chandrasekhar. Efficient gan-based anomaly detection. *arXiv preprint arXiv:1802.06222*, 2018.

[Zenati *et al.*, 2018b] Houssam Zenati, Manon Romain, Chuan-Sheng Foo, Bruno Lecouat, and Vijay Chandrasekhar. Adversarially learned anomaly detection. In *ICDM*, pages 727–736, 2018.

[Zhai *et al.*, 2016] Shuangfei Zhai, Yu Cheng, Weining Lu, and Zhongfei Zhang. Deep structured energy based models for anomaly detection. In *ICML*, pages 1100–1109, 2016.

[Zong *et al.*, 2018] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*, 2018.