Dual-MGAN: An Efficient Approach for Semi-supervised Outlier Detection with Few Identified Anomalies

ZHE LI, Beijing University of Civil Engineering and Architecture, China
CHUNHUA SUN, CHUNLI LIU, XIAYU CHEN, MENG WANG, and YEZHENG LIU*, Hefei University of Technology, China

Outlier detection is an important task in data mining, and many technologies for it have been explored in various applications. However, owing to the default assumption that outliers are not concentrated, unsupervised outlier detection may not correctly identify group anomalies with higher levels of density. Although high detection rates and optimal parameters can usually be achieved by using supervised outlier detection, obtaining a sufficient number of correct labels is a time-consuming task. To solve these problems, we focus on semi-supervised outlier detection with few identified anomalies and a large amount of unlabeled data. The task of semi-supervised outlier detection is first decomposed into the detection of discrete anomalies and that of partially identified group anomalies, and a distribution construction sub-module and a data augmentation sub-module are then proposed to identify them, respectively. In this way, the dual multiple generative adversarial networks (Dual-MGAN) that combines the two sub-modules can identify discrete as well as partially identified group anomalies. In addition, in view of the difficulty of determining the stop node of training, two evaluation indicators are introduced to evaluate the training status of the sub-GANs. Extensive experiments on synthetic and real-world data show that the proposed Dual-MGAN can significantly improve the accuracy of outlier detection, and the proposed evaluation indicators can reflect the training status of the sub-GANs.

CCS Concepts: • **Information systems** → *Data mining*; • **Computing methodologies** → Anomaly detection.

Additional Key Words and Phrases: Discrete anomalies, partially identified group anomalies, distribution construction, data augmentation

1 INTRODUCTION

Outliers refer to observations that have significantly different characteristics from a majority of the other data. These observations are so unique as to arouse the suspicion that they were generated owing to illegal acts or undetected errors. To reveal the critical information in them, many outlier detection technologies have been studied and used in various applications, such as fraud detection in credit card transactions [8, 11, 39] or taxes [17, 19], identifying false information on e-commerce platforms [12, 30] or social media [23, 40, 42], intrusion detection in network service requests [35, 38], and detecting abnormal trajectories during traffic monitoring [4, 34].

*Yezheng Liu is the corresponding author. This work is supported by the Major Program of the National Natural Science Foundation of China (91846201), the National Natural Science Foundation of China (72071069, 71802068, 71801069), the BUCEA Young Scholar Research Capability Improvement Plan under Grant X21080 (07080921008) and National Engineering Laboratory for Big Data Distribution and Exchange Technologies (W2021JSZX0052).

Authors' addresses: Zhe Li, lizhe@bucea.edu.cn, Beijing University of Civil Engineering and Architecture, 1 Zhanlanguan Road, Xicheng District, Beijing, China, 100032; Chunhua Sun, Chunli Liu, Xiayu Chen, {sunchunhua,liuchunli,xychen}@hfut.edu.cn; Meng Wang, eric.mengwang@gmail.com; Yezheng Liu, liuyezheng@hfut.edu.cn, Hefei University of Technology, 193 Tunxi Road, Hefei, Anhui, China, 230041.

1556-4681/2022/3-ART $15.00
https://doi.org/10.1145/3522690

In general, prevalent methods can be divided into three categories according to the availability of the data labels: unsupervised, supervised, and semi-supervised outlier detection. Unsupervised algorithms are among the most widely studied because they do not require additional labels or prior information. Including statistical-based [53, 61], cluster-based [21, 33, 47, 55], OC-SVM-based [13, 14, 54], proximity-based [3, 7, 36, 44], reconstruction-based [32, 45, 59, 60], and other methods. They explicitly or implicitly assume that outliers are not as concentrated as normal data [48]. Thus, discrete anomalies can be detected. However, in many cases, multiple common anomalies (e.g., DoS attacks) may be generated by the same mechanism, which are referred to as group anomalies in this article. They become increasingly concentrated such that unsupervised methods of outlier detection may identify these group anomalies incorrectly as normal data. Moreover, the selection of models and parameters is a considerable challenge for unsupervised methods without the help of prior knowledge. Supervised outlier detection can be regarded as a special classification with unbalanced data. Many classification algorithms and processing techniques for unbalanced data have been used as supervised outlier detectors [15, 23, 25, 26, 51]. Because the labels are complete and correct, high detection rates and optimal parameters can usually be obtained [57]. However, obtaining a sufficient number of correct labels is a time-consuming task, such that the conditions for supervised outlier detection are usually not satisfied in practice. In addition, detection models trained on fully labeled data have considerable uncertainty when dealing with emerging anomalies.

To address the above issues, some semi-supervised outlier detection algorithms have been proposed. According to specific detection scenarios, they can further be divided into three categories: one-class learning with only normal examples, semi-supervised outlier detection with a small amount of labeled data, and semi-supervised outlier detection with few identified anomalies [1]. Because one-class learning is only slightly different from unsupervised outlier detection, it is the most common semi-supervised outlier detection [13, 14, 27, 54]. However, considerable time may be spent on verifying the normal data, and even a large number of normal labels may not provide enough information for model training. As for semi-supervised outlier detection with a small amount of labeled data [16, 50, 52] or few identified anomalies [17, 41, 57], the conditions for the latter are easier to meet and more common than those for the former, and few identified anomalies have been able to yield valuable information for training the detection models. Specifically, despite their insufficient capacity to label large amounts of data, few abnormal behaviors that have triggered an alarm can be collected easily in many applications. Examples include DoS attacks that have crashed the system and insurance applications that have been proven to be fraudulent. These application-related abnormal behaviors usually represent specific abnormal patterns, and behaviors similar to these identified abnormal behaviors are more likely to be abnormal [1, 57]. Therefore, in addition to their own labels, these identified anomalies may also provide a priori information for other samples that have the same generation mechanism. If this information is fully used, partially identified group anomalies can also be detected. Moreover, few anomalies can provide valuable guidance for the selection of models and parameters, where this is a significant advantage over unsupervised outlier detection. Thus, this paper focuses on this special form of outlier detection, in the hope of using few identified anomalies and a large amount of unlabeled data to detect discrete as well as partially identified group anomalies.

The initial model for semi-supervised outlier detection with few identified anomalies [9, 10] first extracts reliable normal examples through a heuristic method, where this is consistent with the first step of PU-learning. A modified model of outlier detection is then trained on the new tagged data to identify the other anomalies. However, because the outliers are usually discrete or belong to different clusters, the extracted samples that are significantly different from the identified anomalies are not necessarily normal. Thus, the potential information in the identified anomalies is not used effectively and erroneous information may be introduced to the new tagged dataset. Therefore, to augment the use of known information and reduce the introduction of errors, several soft versions of the above strategy have been established. For example, the LBS-SVDD [28] assigns abnormal likelihood value to each sample based on the proportion of anomalies in its neighbors, while the ADOA [57] attaches a weight to each instance according to its own isolation [29] and its similarity to the identified anomalies.

However, the calculation of neighbors in the LBS-SVDD usually incurs a high computational cost, whereas the classification models trained using the ADOA may have considerable uncertainty when dealing with emerging anomalies. Thus, although some studies have examined this special form of semi-supervised outlier detection, algorithms for this scenario require further research.

To ensure that partially identified group anomalies can be detected along with discrete anomalies, we propose a model containing two sets of multiple generative adversarial networks for semi-supervised outlier detection with few identified anomalies. We call it the dual multiple generative adversarial networks (Dual-MGAN). The main contributions of this work are as follows:

First, we state the problem of semi-supervised outlier detection with few identified anomalies, and decompose it into two sub-tasks. In contrast to unsupervised outlier detection, this detection scenario contains a handful of identified anomalies. Since samples similar to the identified anomalies are more likely to be abnormal [1, 57]. These identified anomalies not only provide their own labels, but also contain a priori information on other samples with the same mechanism of generation. Therefore, in addition to identifying discrete anomalies, partially identified group anomalies should also be detected.

Second, we propose two sub-modules for the two sub-tasks: multiple generative adversarial active learning (MGAAL) and multiple generative adversarial over-sampling (MGAOS). To detect discrete anomalies, we adopt the strategy of reference distribution construction, which approaches outlier detection as a classification problem by constructing a reasonable reference distribution [31]. And in view of the problem encountered by multiple-objective generative adversarial active learning (MOGAAL) when dealing with a relatively discrete data distribution, we replace multiple-objective sub-generators with multiple sub-GANs (i.e., MGAAL) to directly learn the generation mechanism of the data. In this way, even for discrete data structures, MGAAL can construct a reasonable reference distribution to ensure that discrete anomalies are identified from concentrated normal data. To detect partially identified group anomalies, we adopt the strategy of data augmentation, which can utilize the potential information in the identified anomalies without calculating the degree of abnormality of each instance. And considering that the outliers are usually discrete or belong to different clusters, multiple sub-GANs (i.e., MGAOS) are used to increase the size of the minority class to prevent a single GAN from falling into the mode collapsing problem. In this way, even if only a few anomalies are identified, MGAOS can augment the entire minority class to ensure that both the identified anomalies and the partially identified group anomalies are detected.

Third, we propose the semi-supervised Dual-MGAN that combines MGAAL and MGAOS. Although the goals and principles of the two sub-modules are different, their network structures and execution processes are basically the same. Therefore, they can be easily merged by integrating all potential outliers generated by different sub-GANs. Thus, to separate all the generated data and identified anomalies from unlabeled data, the detector in Dual-MGAN can accurately identify discrete anomalies and detect partially identified group anomalies. Moreover, considering the difficulty of evaluating the training status of the sub-GANs, two evaluation indicators are introduced to make the detection process more reliable.

Finally, extensive experiments were conducted to evaluate the proposed Dual-MGAN. To investigate its characteristics and overall performance, the Dual-MGAN was compared with ten representative outlier detection algorithms (including the two proposed sub-modules) on both synthetic and real-world datasets. The results show that our proposed approach can usually improve the accuracy of outlier detection by simultaneously using the potential information and constructing the normal patterns. In addition, the indicators to assess the status of training and the computational complexity of different algorithms are discussed.

The remainder of this article is structured as follows: Section 2 briefly reviews common methods of outlier detection and focuses on semi-supervised outlier detection with a limited number of labels. Section 3 introduces the detection principle and model details of our previously proposed MOGAAL, which is necessary to understand the method presented in this article. Section 4 formally presents the task of semi-supervised outlier detection

with few identified anomalies, and the proposed semi-supervised algorithm Dual-MGAN is described in detail. Section 5 reports experiments to assess the performance of Dual-MGAN and discusses the effectiveness of the evaluation indicators. Section 6 offers the conclusions of this article and future directions of research in the area.

2 RELATED WORK

A comprehensive overview of outlier detection algorithms for different kinds of data and applications has been provided in the literature [1]. Here, we briefly discuss common methods of outlier detection (including unsupervised and supervised approaches), and then focus on semi-supervised outlier detection with a limited number of labels, which is most relevant to our research. Finally, we review GAN-based outlier detection algorithms in Section 2.3.

2.1 Common Methods of Outlier Detection

Unsupervised methods of outlier detection have been studied widely because they require no additional labels. Representative algorithms include proximity- [3, 7, 36, 44], statistics- [53, 61], cluster- [21, 33, 47, 55], OC-SVM- [13, 14, 54], and reconstruction-based [32, 45, 59, 60] models. Proximity-based models assume that outliers are points that are far from the other data, and can be identified by measuring their distance or density. By contrast, the other models assume that outliers are observations that deviate significantly from the normal profiles, and identify them by creating a model for a majority of the data. However, these algorithms are based on the assumption that outliers are not as concentrated as the normal data such that group anomalies cannot be correctly detected. Most of them also require model-related assumptions or parameters in advance, which is a daunting challenge for unsupervised methods without the help of prior knowledge.

Supervised outlier detection can be regarded as a special classification problem and many classification algorithms have been applied in this case. However, in most applications, the outliers are far fewer than the normal data such that the direct use of off-the-shelf classifiers and evaluation metrics may yield biased results. Hence, cost-sensitive learning [46, 49] and adaptive re-sampling [15, 23, 25, 26, 51] are incorporated into the classification process. Cost-sensitive learning increases the cost of incorrect classification of outliers by weighing the errors in classification, whereas adaptive re-sampling increases the relative proportion of the minority class through under-sampling or over-sampling. Supervised algorithms usually achieve suitable parameters and high detection rates because the labels are complete during training. The challenge, however, is to obtain a sufficient number of labels, which is a time-consuming and expensive task. Moreover, a detection model trained on fully labeled data has significant uncertainty in dealing with emerging anomalies.

2.2 Semi-supervised Methods of Outlier Detection

According to the availability of the data labels, semi-supervised outlier detection can be divided into three categories: one-class learning with only normal examples, semi-supervised outlier detection with a small amount of labeled data, and semi-supervised outlier detection with few identified anomalies. One-class learning is only slightly different from unsupervised outlier detection, and most of approaches to unsupervised detection (e.g., OC-SVM- [13, 14, 27] and reconstruction-based [32, 45, 59, 60]) can be used in this case [1]. The normal profiles established on the one-class dataset tend to be more robust due to the absence of additional interference from anomalies. However, considerable time may be needed to verify the collected samples to ensure that the training data contain only normal data. Semi-supervised outlier detection with a small amount of labeled data usually optimizes the model of outlier detection (e.g., k-means [16] and fuzzy rough k-means [52]) with the guarantee that labels of the data remain unchanged. Compared with unsupervised outlier detection, it improves performance through a small amount of labeled data. However, the potential information in the labeled examples is not

used effectively, and normal examples in the labeled data may still require additional confirmation because of undetected anomalies.

Compared with the above, semi-supervised outlier detection with few identified anomalies is more common because few abnormal behaviors that have triggered alarms can be easily collected in applications. The initial model [9, 10] first extracts reliable normal examples through a heuristic method, and then trains a semi-supervised model as described above. But because the outliers are usually discrete or belong to different clusters, the extracted samples that are far from the identified anomalies are not necessarily normal. As a result, the potential information in the identified anomalies is not used fully, and erroneous information may also be introduced to the new dataset. To solve this problem, several soft versions of the above methods have been established. For example, before training the detection model, the LBS-SVDD [28] evaluates the probability of abnormality of each sample based on the proportion of anomalies in its neighborhood, while the ADOA [57] assigns likelihood values to all samples according to their isolation and similarity to the identified anomalies. This enhances the use of known information while reducing errors. However, the calculation of probability in the LBS-SVDD incurs a high computational cost, whereas classification models trained in the ADOA may have considerable uncertainty when dealing with emerging anomalies. Therefore, research on semi-supervised outlier detection with few identified anomalies remains inadequate.

2.3 GAN-based Methods of Outlier Detection

Generative adversarial networks (GAN) [18] is an adversarial representation learning model. The generator attempts to generate samples that are similar to the real data, while the discriminator attempts to identify the differences between the real and the generated data. In order to achieve their own goals, the parameters of the two adversarial components are constantly adjusted. Thus, after a sufficient number of iterations, the generator can learn the generation mechanism of the real data, while the discriminator can identify subtle differences between different types of data. Owing to the powerful representation capabilities, GAN and improved models of it have been applied to a variety of applications. For unsupervised outlier detection and semi-supervised outlier detection with only normal examples, GAN-based reconstruction models and generation models have been studied. GAN-based reconstruction models usually learn the generation mechanism of the normal data by training a regular GAN [45] or a combination of the GAN and other neural networks [2, 32, 43, 56], and then measure the degree of abnormality of the examples based on the reconstruction or the discriminator loss. Moreover, to prevent slight anomalies from being reconstructed, Bian et al. [5] performed active negative training to limit the capability of the generator. GAN-based generation models use the GAN to generate informative potential outliers [31] or infrequent normal samples [24] such that subsequent detectors can describe a correct division boundary. For supervised outlier detection, GAN [15, 25, 37] is often used to synthesize examples of the minority class to balance the relative proportion between the classes. Besides, Zheng et al. [58] took advantage of an adversarial deep denoising autoencoder to extract latent representations of labeled transactions to significantly improve the accuracy of fraud detection. However, few GAN-based studies have focused on semi-supervised outlier detection with few identified anomalies. Although Kimura et al. [20] used both noisy normal and abnormal images for visual inspection, the main purpose was to eliminate the impact of abnormal pixels during the reconstruction process, which differs from our model here.

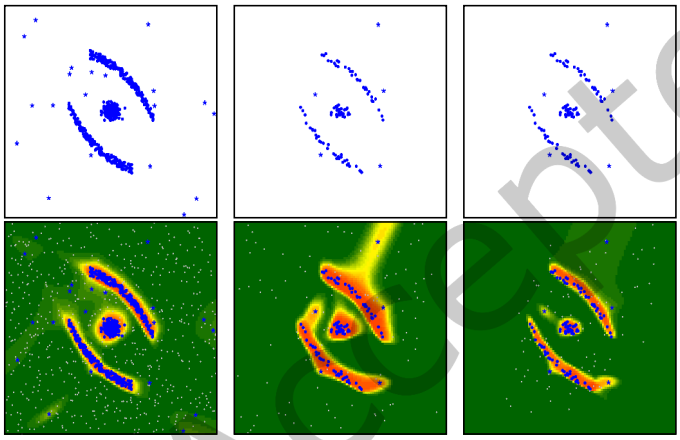
3 BACKGROUND ON MOGAAL

Due to the assumption that the outlier is not concentrated [48], unsupervised outlier detection can be regarded as a density-level detection process. Unlike prevalent model-based or proximity-based methods, artificially generating potential outliers (AGPO)-based algorithms approach density-level detection as a classification problem. First, for a given dataset $X = \{x_1, ..., x_n\}$ (shown with blue dots and stars in Fig. 1(a)), the initial AGPO randomly generates

n data points x' as potential outliers (shown with gray dots in Fig. 1(a)) that are used to construct a uniform reference distribution μ for X. A classifier C is then trained on the new dataset to separate potential outliers x' from the original data x. To minimize the loss function $\mathcal{L}_C$,

$$\mathcal{L}_C = -\sum_{i=1}^{n}[\log(C(x_i)) + \log(1 - C(x_i'))] \tag{1}$$

the classifier C should assign a higher value to the original data x that have a higher relative density $\rho'(x) = \frac{\rho(x)}{\rho(x')}$, and a lower one in the opposite case. Thus, because the absolute density $\rho(x')$ of the potential outliers is equal everywhere, the optimized classifier C can assign higher values to the original data with higher absolute density $\rho(x)$, and can describe a correct division boundary to separate non-concentrated outliers $\{x_i|y_i = 0\}$ from concentrated normal data $\{x_i|y_i = 1\}$ (as shown in Fig. 1(a)).



(a) AGPO on Low. (b) AGPO on High. (c) MOGAAL on High.

Fig. 1. The detection performance of the initial AGPO and MOGAAL. Normal points, outliers, and potential outliers are shown with blue dots, blue stars, and gray dots, respectively. High-dimensional data are presented as cross-sectional data, and data points closer to the green area are more likely to be outliers.

However, as the number of dimensions increases, the data becomes sparser. A limited number of potential outliers generated by the initial AGPO are randomly scattered throughout the sample space, so that their absolute density approaches zero (i.e., $\rho(x') \Rightarrow 0$). The original data with low absolute density $\rho(x)$ may also have a high relative density $\rho'(x)$. As a result, the classifier C may describe an incorrect division boundary to detect non-concentrated outliers as normal data (as shown in Fig. 1(b)). Therefore, MOGAAL was proposed to generate informative potential outliers x' and construct a reasonable reference distribution μ to ensure that the relative density of the concentrated normal data is higher than that of the non-concentrated outliers (i.e., $\rho'(\{x_i|y_i = 1\}) > \rho'(\{x_i|y_i = 0\})$).

The network structure and detection process of MOGAAL is illustrated in Fig. 2. It consists of k sub-generators $G_{1:k}$ and a discriminator D. The sub-generators $G_{1:k}$ are used to generate informative potential outliers x' to construct a reasonable reference distribution μ; while the discriminator D is used to describe a correct division boundary that encloses the concentrated normal data, such as the classifier $C(x)$. Specifically, given that samples x with similar output values $D(x)$ are more likely to be similar, MOGAAL first divides the original dataset X into k equal subsets $X_{1:k}$ based on their similar output values. A dynamic game is then executed between each sub-generator G_i and the discriminator D. Sub-generator G_j attempts to learn the generation mechanism of X_j by

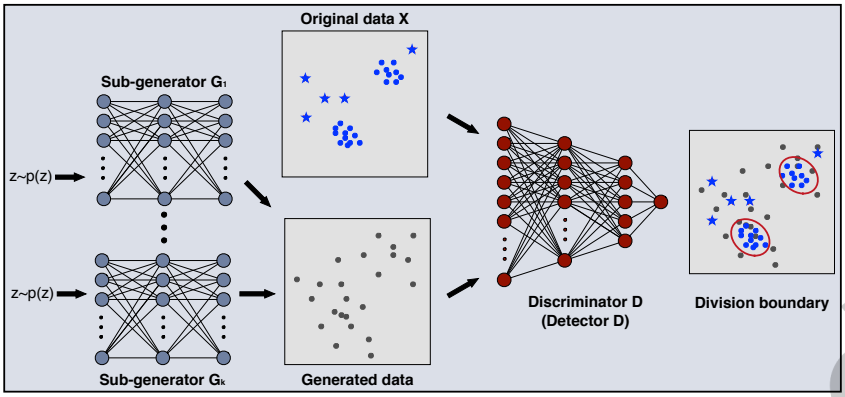


Fig. 2. Details of the model of MOGAAL. Through a dynamic game between $G_{1:k}$ and D, the specific sub-generator G_j generates informative potential outliers that occur inside or close to the specific subset X_j, and the discriminator D describes a division boundary to enclose the original, concentrated data.

making the generated data $G_j(z)$ output similar values to $D(x|x \in X_j)$. The discriminator D attempts to identify the generated data $G_j(z)$ from the original data x. The overall optimization framework is formulated as follows:

$$\min_{\theta_{g_j}} V_{G_j} = -\sum_{i=1}^{n_j}[T_j \log(D(G_j(z_j^{(i)}))) + (1 - T_j)\log(1 - D(G_j(z_j^{(i)})))] \tag{2}$$

$$\max_{\theta_d} V_D = \sum_{x \in X} \log(D(x)) + \sum_{j=1}^{k}\sum_{i=1}^{n'_j} \log(1 - D(G_j(z_j^{(i)}))) \tag{3}$$

where T_j is a representative statistical value of $D(x|x \in X_j)$, n_j is the number of samples in the j-th subset X_j, and n'_j is the number of potential outliers generated by G_j. As the game progresses, sub-generator G_j gradually learns the generation mechanism of X_j and generates an increasing number of informative potential outliers x' that occur inside or close to the original dataset X_j. In this way, even in a high-dimensional space, the absolute density $\rho(x')$ of the potential outliers x' generated by G_j around X_j does not approach zero (i.e., $\rho(x') > 0$). Coupled with the control of the number n'_j of potential outliers generated by G_j, a reasonable reference distribution μ (i.e., $\rho'(\{x_i|y_i = 1\}) > \rho'(\{x_i|y_i = 0\})$) can be constructed by integrating different numbers of informative potential outliers $G_{1:k}(z)$. As a result, the discriminator D can describe a correct division boundary to separate non-concentrated outliers from concentrated normal data (as shown in Fig. 1(c) and Fig. 2). A more complete explanation is available in the literature [31].

4 SEMI-SUPERVISED OUTLIER DETECTION WITH FEW IDENTIFIED ANOMALIES

The most pressing problem with unsupervised outlier detection (including MOGAAL) is that it cannot detect group anomalies in the absence of additional information. Fortunately, although a sufficient number of labels are difficult to obtain, few common abnormal behaviors (e.g., DoS and DDoS attacks) that have trigger alarms can be collected easily in most applications. These application-related abnormal behaviors usually represent specific abnormal patterns [1]. In this way, behaviors similar to these identified abnormal behavior are more likely to be abnormal [57], which is also the default assumption followed by most semi-supervised outlier detection techniques (e.g., misuse-based intrusion detection [22]). Therefore, these identified anomalies not only contain their own labels, but also potentially provide a priori information for other samples. That is, samples with the same generation mechanism as identified anomalies have a high probability of being anomalies. If this information

is fully used, partially identified group anomalies can be detected accurately along with the discrete anomalies. However, current unsupervised algorithms cannot use this potential information, and semi-supervised algorithms in this scenario still require further research. Therefore, we first state the problem of semi-supervised outlier detection with few identified anomalies and a large amount of unlabeled data, and then decompose it into two sub-tasks: the detection of discrete anomalies and the detection of partially identified group anomalies. Following this, the distribution construction sub-module and the data augmentation sub-module (i.e., MGAAL and MGAOS) are proposed for the two sub-tasks, respectively. In this way, the semi-supervised Dual-MGAN that combines MGAAL and MGAOS can detect discrete anomalies and partially identified group anomalies, simultaneously. In addition, considering the difficulty of finding the stop node in training, two evaluation indicators are introduced to the Dual-MGAN to make the detection process more intelligent.

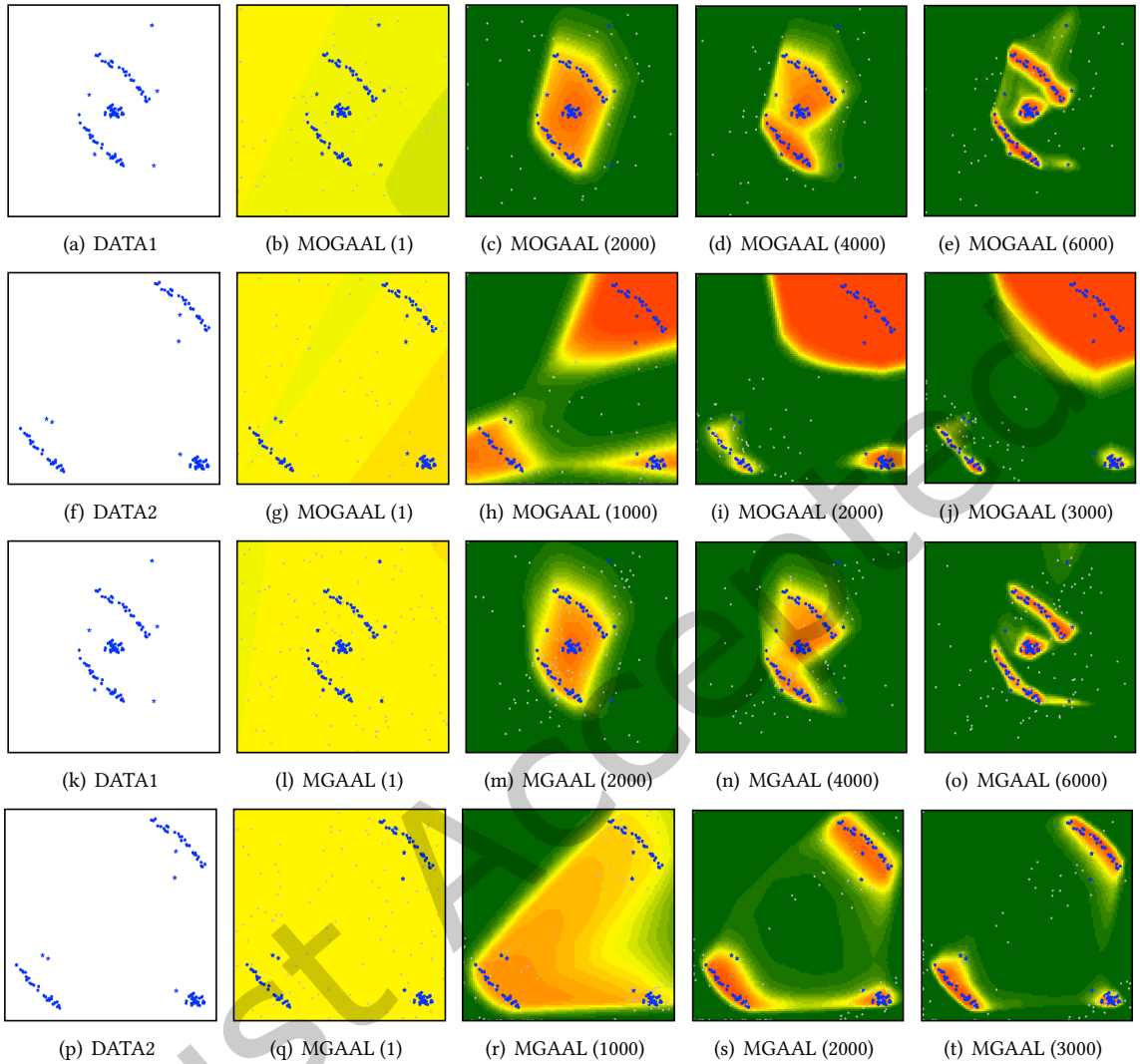
4.1 Problem Statement

Consider a dataset $X = \{x_{u1}, \ldots, x_{un_u}, x_{a1}, \ldots, x_{an_a}\}$ with n_u unlabeled samples $X_u = \{x_{ui} | y_{ui} = 1 \text{ or } 0\}$ and n_a identified anomalies $X_a = \{x_{ai} | y_{ai} = 0\}$, where $x_i \in \mathbb{R}^d$ represents a data point, $y_i \in \{0, 1\}$ represents its label, and $n_u \gg n_a$. The purpose of semi-supervised outlier detection is to identify a scoring function $\zeta(x) \in [0, 1]$ that can assign a higher value (close to 1) to the normal data and a lower value (close to 0) to the outlier, where this is consistent with unsupervised outlier detection. But the difference is that the training dataset X for semi-supervised outlier detection still contains a handful of identified anomalies X_a. Since samples similar to the identified anomalies are more likely to be abnormal [1, 57]. These anomalies not only provide their own labels, but can also potentially provide a priori information for other samples. Therefore, in addition to identifying discrete anomalies, samples with the same generation mechanism as the identified anomalies should also be detected as anomalies. That is, the scoring function $\zeta(x)$ should accomplish the following two sub-tasks: (i) Based on the default assumption that the outliers are not as concentrated as the normal data, the scoring function should assign higher values to samples with higher density levels and output lower values for the discrete data. (ii) Assuming that samples with the same generation mechanism as identified anomalies are more likely to be outliers, the scoring function should output a value close to 0 for them and the identified anomalies.

4.2 Multiple Generative Adversarial Active Learning (MGAAL)

Aiming at the first sub-task, namely, the detection of discrete anomalies, we adopt the strategy of reference distribution construction. This is because it can handle data clusters of varying shapes, and the trained discriminator is suitable for online real-time detection. However, there is still a problem in constructing the reference distribution in MOGAAL, which assumes that samples with similar output values are more likely to be similar. Specifically, the output of the discriminator reflects the relative density of the data. When the distribution of the data is concentrated, the under-fitting discriminator usually outputs a higher value for the area where most of the data are located (as shown in Fig. 3(c)). Data with higher output values are more similar in the sample space. In this case, the sub-generators in MOGAAL can construct a reasonable reference distribution μ (shown with gray dots in Figs. 3(c)-3(e)) by making the generated samples output similar values. Eventually, the discriminator can describe a correct division boundary (as shown in Fig. 3(e)).

However, when the distribution of the data is discrete, those with higher output values may be very different, such as the three clusters shown in Fig. 3(h). In this case, the data in the same subset divided according to similar output values may be very different, and the potential outliers generated by similar output values are not necessarily similar to all data in the specific subset. As shown in Figs. 3(h)-3(j), although potential outliers with similar output values occur inside or close to part of the data (such as the cluster in the lower-left corner), they may also be significantly different from the other data (such as the cluster in the upper-right corner). Because there are no potential outliers around these neglected data, they all have a high relative density. As a result, the



(a) DATA1 (b) MOGAAL (1) (c) MOGAAL (2000) (d) MOGAAL (4000) (e) MOGAAL (6000)

(f) DATA2 (g) MOGAAL (1) (h) MOGAAL (1000) (i) MOGAAL (2000) (j) MOGAAL (3000)

(k) DATA1 (l) MGAAL (1) (m) MGAAL (2000) (n) MGAAL (4000) (o) MGAAL (6000)

(p) DATA2 (q) MGAAL (1) (r) MGAAL (1000) (s) MGAAL (2000) (t) MGAAL (3000)

Fig. 3. The detection performance of MOGAAL and MGAAL on concentrated and discrete data. The distribution of the "DATA1" is concentrated and that of the "DATA2" is discrete. The value in parentheses is the number of iterations.

discriminator describe an incorrect boundary such that discrete anomalies contained in these neglected data are identified as normal (as shown in Fig. 3(j)). To solve this problem, we propose an improved MGAAL that replaces multiple-objective sub-generators with multiple sub-GANs to directly learn the generation mechanism of the data.

The network structure of MGAAL is shown in the left part of Fig. 4. It consists of a clustering module and k_u sub-GANs $GAN_{u1:uk_u}$. The clustering module is used to divide the unlabeled data into different subsets and the sub-GANs are used to construct a reasonable reference distribution for the unlabeled data. Moreover, an artificial neural network (i.e., the detector D) is introduced to the training process of MGAAL to describe the division boundary, where this is consistent with the discriminator $D(x)$ in MOGAAL. Specifically, the clustering module

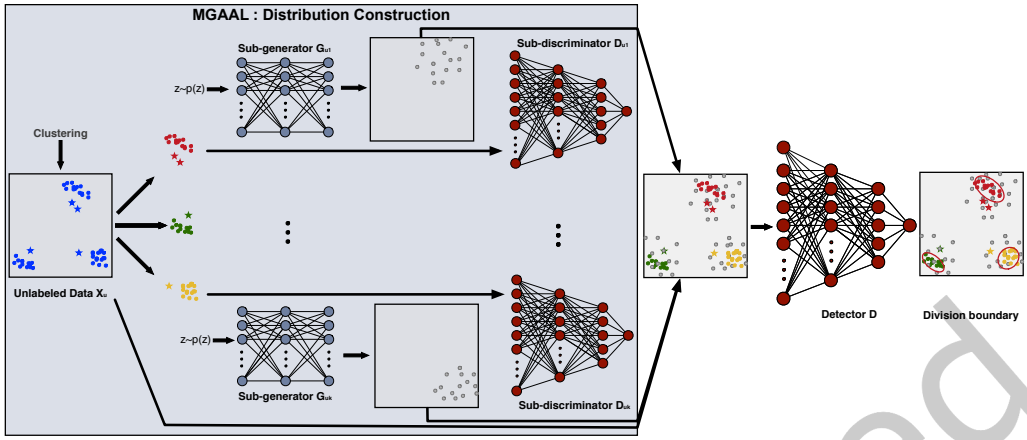


Fig. 4. Details of the model of MGAAL. As MGAAL is mainly used to detect discrete anomalies, group anomalies are not included in the dataset.

first divides the unlabeled dataset X_u into k_u subsets $X_{u1:uk_u}$ based on a similarity measure rather than similar output values. In this way, even in the case of discrete data structures, MGAAL can ensure that the data x_{ui} in the same subset X_{uj} are more similar to each other than otherwise (shown with dots and stars of different colors in Fig. 4). Then, for each subset X_{uj}, MGAAL constructs a specific sub-GAN GAN_{uj}, which can learn the generation mechanism of the data $x_{ui} \in X_{uj}$ through a mini-max game between the sub-generator G_{uj} and sub-discriminator D_{uj}:

$$\min_{\theta_{g_{uj}}} \max_{\theta_{d_{uj}}} V(D_{uj}, G_{uj}) = \sum_{x \in X_{uj}} \log(D_{uj}(x)) + \sum_{i=1}^{n_{uj}} \log(1 - D_{uj}(G_{uj}(z_{uj}^{(i)}))) \tag{4}$$

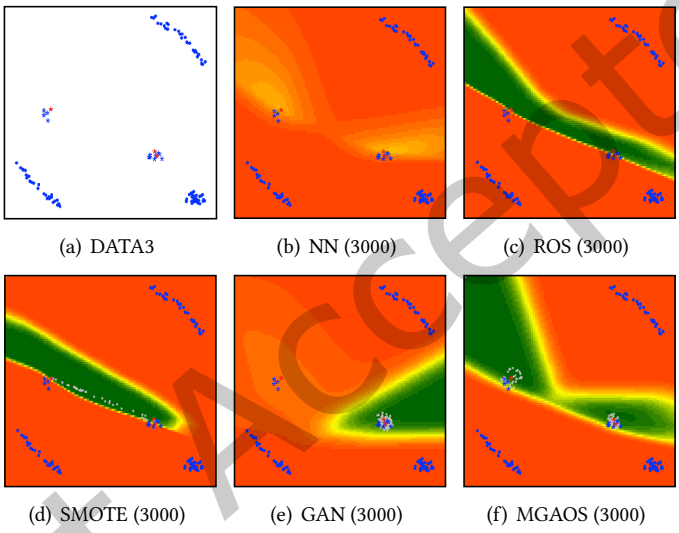
where n_{uj} represents the number of samples in the j-th subset X_{uj}. Because the unlabeled data have been decomposed into several separate clusters, the single sub-GAN can generate informative potential outliers that occur inside or close to all the data in X_{uj}. When all sub-GANs have learned the generation mechanism of the data, the integrated potential outliers $G_{u1:uk_u}(z)$ generated by different sub-generators $G_{u1:uk_u}$ can be used to construct a reasonable reference distribution for the entire unlabeled dataset (shown with gray dots in Fig. 4). The detector D, in order to separate potential outliers $G_{u1:uk_u}(z)$ from unlabeled data X_u, gradually describes a correct division boundary that encloses the concentrated normal data (as shown in Figs. 3(q)-3(t) and Fig.4):

$$\min_{\theta_d} V_D = -[\sum_{x \in X_u} \log(D(x)) + \sum_{j=1}^{k_u} \sum_{i=1}^{n'_{uj}} \log(1 - D(G_{uj}(z_{uj}^{(i)})))] \tag{5}$$

where n'_{uj} represents the number of potential outliers generated by G_{uj}. In contrast to MOGAAL, MGAAL generates the same number of potential outliers for different subsets (i.e., $n'_{uj} = \lceil \frac{n_u}{k_u} \rceil$). Because each subset partitioned by the clustering module contains a different number of samples, the concentrated data are usually divided into large subsets. Compared with MOGAAL, MGAAL can ensure the detection of discrete anomalies, even when the distribution of the data is discrete, by constructing a reasonable reference distribution (as shown in Figs. 3(j) and 3(t)). In addition, MOGAAL needs to calculate the target value in each game while MGAAL simply divides the data before the game starts. This enables MGAAL to have a lower computational complexity on a large dataset.

4.3 Multiple Generative Adversarial Over-sampling (MGAOS)

Aiming at the second sub-task, namely, the detection of partially identified group anomalies, we use data augmentation. Specifically, because only a handful of anomalies X_a can be identified (shown with red stars in Fig. 5(a)), a large number of unidentified anomalies (shown with blue stars in Fig. 5(a)) persist in the unlabeled dataset X_u. If only the identified labels are introduced to the detection process. To achieve higher accuracy, the detector may detect the identified anomalies as normal (as shown in Fig. 5(b)). This is similar to the problem in classifying unbalanced data. Therefore, it is necessary to perform additional processing to ensure that the identified and unidentified anomalies are detected simultaneously. The clear solution is to find the unidentified anomalies by calculating the similarity between the data and the identified anomalies. However, the conversion of similarity to abnormal value varies greatly in different scenarios, and the calculation of similarity is likely to be affected by irrelevant variables. Therefore, we focus on data augmentation, which is among the most commonly used processing technologies for unbalanced data.



(a) DATA3 (b) NN (3000) (c) ROS (3000)

(d) SMOTE (3000) (e) GAN (3000) (f) MGAOS (3000)

Fig. 5. The performance of different data augmentation methods. "DATA3" consists of three sets of normal data and two sets of partially identified group anomalies.

Random over-sampling (ROS) involves repeatedly sampling the minority data to increase the size of the minority class, while the synthetic minority over-sampling technique (SMOTE) synthesizes new samples between the minority data and their nearest neighbors. However, repeated sampling may cause the problem of over-fitting (as shown in Fig. 5(c)), and synthesizing new samples between the minority data and their nearest neighbors may lead to over-generalization (as shown in Fig. 5(d)). To avoid these issues, some researchers take advantage of the learning ability of GAN to directly generate samples similar to the minority data. But unlike concentrated minority data, outliers are usually discrete or belong to different clusters. A single GAN may learn only the generation mechanism of part of the outliers such that the ignored anomalies are incorrectly detected as normal (as shown in Fig. 5(e)). This is similar to the problem in MOGAAL. Therefore, inspired by the MGAAL, we propose MGAOS to make full use of the potential information in the identified anomalies.

MGAOS (shown in Fig. 6) consists of a clustering module and k_a sub-GANs $GAN_{a1:ak_a}$; it is similar to MGAAL in this sense. Moreover, an artificial neural network (i.e., the detector D) is introduced to separate the identified anomalies from the unlabeled data. Specifically, the clustering module first divides the identified anomalies

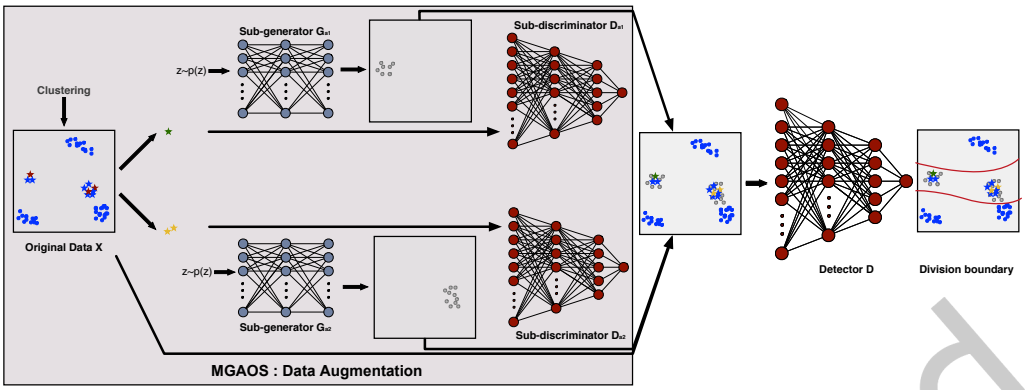


Fig. 6. Details of the model of MGAOS. Because it is mainly used to detect partially identified group anomalies, discrete anomalies are not included in the dataset.

$X_a = \{x_{a1}, \ldots, x_{an_a}\}$ into k_a subsets $X_{a1:ak_a}$ (shown with stars of different colors in Fig. 6). Then, for each subset X_{aj}, MGAOS constructs a specific sub-GAN GAN_{aj} to generate samples similar to the identified anomalies $x_{ai} \in X_{aj}$:

$$\min_{\theta_{g_{aj}}} \max_{\theta_{d_{aj}}} V(D_{aj}, G_{aj}) = \sum_{x \in X_{aj}} \log(D_{aj}(x)) + \sum_{i=1}^{n_{aj}} \log(1 - D_{aj}(G_{aj}(z_{aj}^{(i)}))) \tag{6}$$

where n_{aj} represents the number of samples in the j-th subset X_{aj}. Because the identified anomalies have been decomposed into several separate clusters, the single sub-GAN GAN_{aj} can generate samples similar to all identified anomalies in X_{aj}. When all sub-GANs have learned the generation mechanism of the identified anomalies, the generated samples $G_{a1:ak_a}(z)$ can augment the entire minority class (shown with gray dots in Fig. 6). In this way, to minimize V_D, the detector D outputs values close to 0 for partially identified group anomalies (as shown in Fig.5(f) and Fig.6),

$$\min_{\theta_d} V_D = -[\sum_{x \in X_u} \log(D(x)) + \sum_{x \in X_a} \log(1 - D(x)) + \sum_{j=1}^{k_a} \sum_{i=1}^{n'_{aj}} \log(1 - D(G_{aj}(z_{aj}^{(i)})))] \tag{7}$$

where n'_{aj} represents the number of samples generated by G_{aj}. In order to prevent the detector D from over-fitting or forgetting the identified anomalies, the sub-generators generate multiple samples for subsets containing different numbers of anomalies (e.g., $n'_{aj} = 10 * n_{aj}$).

4.4 Dual Multiple Generative Adversarial Networks (Dual-MGAN)

MGAAL constructs a reasonable reference distribution to ensure that the discrete anomalies can be identified from the concentrated data. MGAOS augments the minority class to ensure that the identified anomalies as well as partially identified group anomalies can be recognized from the unlabeled data. Coupled with the detector D, the two can be used for unsupervised and supervised outlier detection, respectively. However, in case of semi-supervised outlier detection with few identified anomalies, detection using only MGAAL cannot make full use of the potential information to detect group anomalies, whereas outlier detection using only MGAOS cannot establish normal patterns to detect previously unknown anomalies. Therefore, the semi-supervised Dual-MGAN that combines MGAAL and MGAOS is proposed.

The network structure and detection process of Dual-MGAN are shown in Fig. 7. It consists of a distribution construction sub-module, a data augmentation sub-module, and a detector D. In order to detect as many anomalies

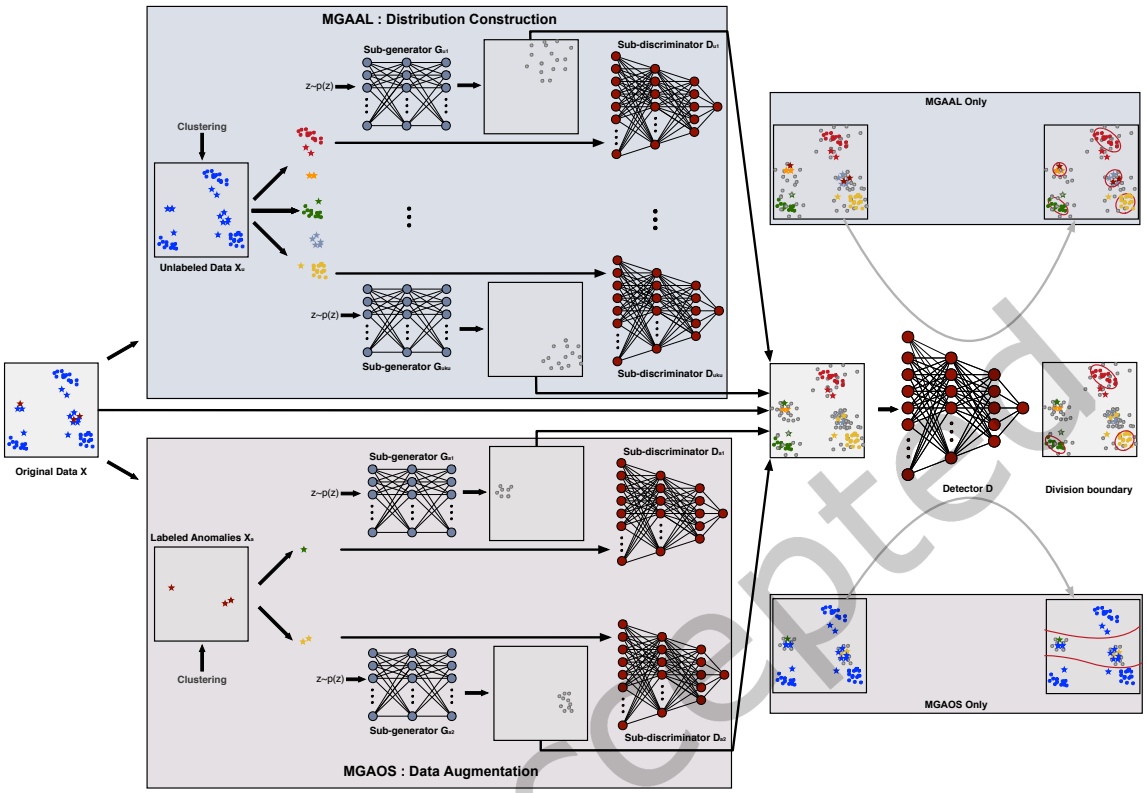


Fig. 7. Details of the model of Dual-MGAN. To match the settings for semi-supervised outlier detection with few identified anomalies, the original data consist of three sets of normal data, two sets of partially identified group anomalies, and four discrete anomalies.

as possible, the MGAOS in Dual-MGAN first increases the size of the minority class by generating multiple samples $G_{a1:ak_a}(z)$ similar to the identified anomalies X_a. Then, MGAAL and the detector D are alternately optimized. MGAAL gradually learns the generation mechanism of the unlabeled data X_u, while the detector D attempts to separate the generated samples (i.e., $G_{a1:ak_a}(z)$ and $G_{u1:uk_u}(z)$) and the identified anomalies X_a from the unlabeled data X_u.

$$\min_{\theta_d} V_D = -[\sum_{x \in X_u} \log(D(x)) + \sum_{x \in X_a} \log(1 - D(x)) + \sum_{j=1}^{k_u} \sum_{i=1}^{n'_{uj}} \log(1 - D(G_{uj}(z_{uj}^{(i)}))) + \sum_{j=1}^{k_a} \sum_{i=1}^{n'_{aj}} \log(1 - D(G_{aj}(z_{aj}^{(i)})))] \tag{8}$$

At the beginning of the iteration, randomly generated potential outliers $G_{u1:uk_u}(z)$ may not provide sufficient information for D. However, when all sub-GANs in MGAAL have learned the mechanism of generation, the integration of the potential outliers $G_{u1:uk_u}(z)$ can provide a reasonable reference distribution for the unlabeled data. Thus, to minimize the optimization function V_D, the detector D not only assigns a higher value (close to 1)

to the concentrated unlabeled data, but also assigns a lower value (close to 0) to discrete anomalies and partially identified group anomalies, which is the scoring function $\zeta(x)$ that we are looking for.

In addition to the general process, two issues must be discussed to ensure more reliable detection. The first is the selection of the final model. In contrast to scenarios of unsupervised outlier detection, the data used to train the detector usually contain few identified anomalies. Except for label-related information, they can also provide valuable guidance for the selection of the model. Therefore, to obtain reliable results, we use the Area Under Curve (AUC) of the output $D(x)$ to measure the detection performance of Dual-MGAN. A higher $AUC(D(x))$ means that the model assigns lower values to the identified anomalies than to other data, and the model corresponding to the highest $AUC(D(x))$ is used as the final model for subsequent detection.

The second issue that needs to be addressed is the evaluation of the training status of the sub-GANs. Because over-training may lead to mode collapse or waste computational resources, all sub-GANs in Dual-MGAN must stop training when they have learned the distribution of the target data. Therefore, the evaluation of the training status has a substantial effect on the results. The original GAN uses the classification error to evaluate the similarity between distributions of the generated data and the real data, that is, the sub-generator has learned the generation mechanism of the target subset when the classification error of the sub-discriminator is close to $-\frac{1}{2}\log(\frac{1}{2})$. However, because the objective function is usually non-convex, exactly the same distributions (i.e., Nash equilibrium) cannot be guaranteed. Our previously proposed MOGAAL uses the trend of generator loss to evaluate the training status of the GAN, that is, the sub-generator has learned the generation mechanism when the downward trend of the generator loss tends to be slow. However, due to fluctuations in loss, accurate assessment of the trend requires human intervention, which significantly increases the uncertainty of the model. In response to this demand, we propose two evaluation indicators for subsets containing different numbers of samples, the nearest neighbor ratio Nnr and the relative distance Rd, to monitor the training status of the sub-GANs.

For the subset X_{uj} containing multiple samples, if the sub-generator has learned the generation mechanism of the target data, the generated data occur inside or close to $x_{ui} \in X_{uj}$. We thus first extract m samples from X_{uj}, and calculate the ratio nnr_{uji} of the generated samples $G_{uj}(z)$ among the s nearest neighbors of each sample $x_{ui} \in X_{uj}$. If nnr_{uji} is greater than a given threshold τ, the sub-generator G_{uj} is considered to have learned the generation mechanism of x_{ui}, and to be able to generate informative potential outliers $G_{uj}(z)$ that are similar to x_{ui}. Then, the ratio Nnr_{uj} of the samples x_{ui} that have been learned in the randomly selected samples is calculated. If Nnr_{uj} is greater than a certain threshold τ', it is concluded that the sub-generator G_{uj} has learned the generation mechanism of the target data X_{uj}. However, for the subset X_{aj} containing only one data point, the calculation of Nnr_{aj} makes no sense. Therefore, we first calculate the distance Dis_{aj} between the generated data $G_{aj}(z)$ and the data point x_{ai}, and then compare Dis_{aj} with the distance dis_{aj} between x_{ai} and its t-th nearest neighbor in the original data X. If the distance Dis_{aj} is less than dis_{aj} (i.e., $Rd_{aj} = 1$), this indicates that the generated data $G_{aj}(z)$ have been able to prevent the data point x_{ai} from being ignored. Moreover, considering the randomness of a single generated data point, we stop training G_{aj} when the frequency of $Rd_{aj} = 1$ is high, or directly collect a certain number of generated data whose Dis_{aj} is less than dis_{aj}. Compared with absolute distance, relative distance (e.g., nearest neighbor ratio and relative distance) can better reflect the value of the information provided by the generated data. Moreover, the final state of the sub-GAN can be easily adjusted by setting different thresholds. The procedure of Dual-MGAN is presented in Algorithm 1, and the code is available at: https://github.com/leibinghe/Dual-MGAN.

5 EXPERIMENTS

Extensive experiments were conducted to assess the proposed method. We report them, and discuss the evaluation indicators, computational complexity, and robustness of Dual-MGAN. Section 5.1 details the experimental

Algorithm 1 Dual-MGAN

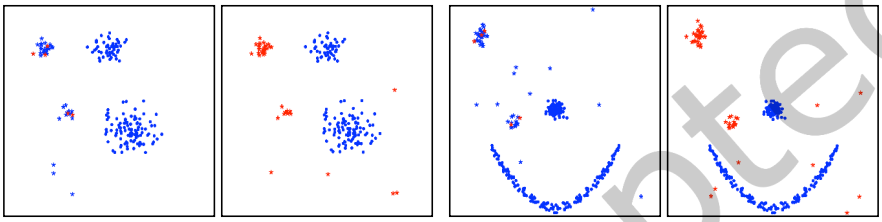
Input: $X = \{X_a, X_u\}$; p_z; τ_a; τ_u; τ'_a; τ'_u; τ''_a; τ''_u
Output: outlier score, $OS(x)$

Divide X_u into k_u subsets $X_{u1:uk_u}$
Divide X_a into k_a subsets $X_{a1:ak_a}$
Initialize $G_{u1:uk_u}$; $D_{u1:uk_u}$; $G_{a1:ak_a}$; $D_{a1:ak_a}$; D
Initialize $n_{u1:uk_u}$; $n_{a1:ak_a}$; $n'_{u1:uk_u}$; $n'_{a1:ak_a}$
Initialize $Nnr_{u1:uk_u} = 0$; $F(Rd_{u1:uk_u}) = 0$; $Nnr_{a1:ak_a} = 0$; $F(Rd_{a1:ak_a}) = 0$; $AUC(D'(X)) = 0$
repeat
  for $j = 1$ to k_a **do**
    if $Nnr_{aj} < \tau'_a$ and $F(Rd_{aj}) < \tau''_a$ **then**
      Sample n_{aj} noises z from p_z
      Update G_{aj} and D_{aj} by optimizing Eq. (6)
      if $n_{aj} > 1$ **then**
        Compute Nnr_{aj} by τ_a, $G_{aj}(z)$ and $x \in X_{aj}$
      if $n_{aj} = 1$ **then**
        Compute $F(Rd_{aj})$ by $G_{aj}(z)$ and $x \in X_{aj}$
until the maximum number of iterations or all indicators are greater than the thresholds
for $j = 1$ to k_a **do**
  Generate n'_{aj} potential outliers $G_{aj}(z)$ to augment the minority class
repeat
  for $j = 1$ to k_u **do**
    if $Nnr_{uj} < \tau'_u$ and $F(Rd_{uj}) < \tau''_u$ **then**
      Sample n_{uj} noises z from p_z
      Update G_{uj} and D_{uj} by optimizing Eq. (4)
      if $n_{uj} > 1$ **then**
        Compute Nnr_{uj} by τ_u, $G_{uj}(z)$ and $x \in X_{uj}$
      if $n_{uj} = 1$ **then**
        Compute $F(Rd_{uj})$ by $G_{uj}(z)$ and $x \in X_{uj}$
    Generate n'_{uj} potential outliers $G_{uj}(z)$
  Update D by optimizing Eq. (8)
  if $AUC(D(X)) > AUC(D'(X))$ **then**
    Save D as D'
until the maximum number of iterations
return $OS(x) = 1 - D'(x)$

settings, including the datasets, evaluation measures, baselines, and parameter settings. To evaluate the detection performance of Dual-MGAN, Section 5.2 compares it with ten detection methods on both synthetic and real-world datasets. Finally, the effectiveness of the two evaluation indicators, the computational complexity of different algorithms, and the robustness of the algorithm are discussed in Sections 5.3, 5.4, and 5.5, respectively.

5.1 Experimental Settings

5.1.1 Datasets. To examine the performance characteristics of different algorithms in detail, we generated two 2D datasets (i.e., Mul-clusters and Mul-shapes) based on the usual assumptions. The training and test data of Mul-clusters (shown in Fig. 8(a)) both contain two sets of normal data that obey two Gaussian distributions, two sets of group anomalies, and several discrete anomalies. The training and test data of Mul-shapes (shown in Fig. 8(b)) contain a set of normal data that obey a Gaussian distribution, a set of normal data that obey a U-shaped distribution, two sets of group anomalies, and several discrete anomalies. The normal data and group anomalies in the test data have the same generation mechanisms as the corresponding training data, whereas the discrete outliers are unidentified or emerging anomalies. Moreover, to match the settings of outlier detection with few identified anomalies, five anomalies were randomly sampled from the training data as identified anomalies (shown with red stars). Compared with Mul-clusters, Mul-shapes can be used to evaluate the ability of outlier detectors to handle data clusters of various shapes.



(a) Mul-clusters Dataset (b) Mul-shapes Dataset

Fig. 8. Two synthetic datasets: (a) shows the training and test data of the Mul-clusters dataset and (b) shows the training and test data of the Mul-shapes dataset.

To obtain an overall assessment of different algorithms, ten real-word datasets that have often been used in the literature on outlier detection were also selected for the experiments. These datasets were first processed as outlier evaluation datasets according to the procedure described in [6]. Then, we divided each dataset into a training and a test dataset in the ratio of 2 : 1. Furthermore, 10% of the anomalies in the training data were randomly selected as identified anomalies to match the settings of few identified anomalies. Detailed information on these datasets is listed in Table 1.

Table 1. Description of the real-world datasets

Dataset	Description (Nor. vs. Out.)	Dim.	Data Size		Train		Test
			Nor.	Ano.	Unl.	Ide.	
Pima	Healthy vs. Diabetes	8	500	268	489	18	261
Stamps	Genuine vs. Forged	9	3409	31	224	3	113
Shuttle	Class "1" vs. Others	9	45586	3511	32494	377	16366
Cover	Class "2" vs. Class "4"	10	283310	2747	190515	184	95349
Vowels	Others vs. Class "1"	12	1406	50	967	4	485
Cardio	Healthy vs. Pathologic	21	1655	176	1209	12	610
WDBC	Benign vs. Malignant	30	357	10	240	1	126
Ionosphere	Class "g" vs. Class "b"	32	225	126	216	9	117
Arrhythmia	Healthy vs. Arrhythmia	259	244	206	290	14	146
Speech	American accent vs. Others	400	3645	61	2452	5	1229

5.1.2 Evaluation Measures. Owing to the inherent imbalance between normal data and outliers, we used three evaluation measures that are less sensitive than other measures to comprehensively evaluate the detection performance of different algorithms. The first is the area under the ROC curve (i.e., AUC), which is used to evaluate the full ranking returned by the detection model. Then, because most applications are interested only in the top-ranked objects, the proportion of true outliers in the top ranks (i.e., precision) was applied to assess the algorithms. To fairly choose the number of objects that are ranked high, we set them according to the ground truth in different datasets. In addition, considering that the number of outliers is not always known in many cases, the average precision over a wide range of possible choices (i.e., AP) was introduced to evaluate precision more comprehensively. Although the methods used to calculate the three evaluation measures are different, their values are all between 0 and 1. The higher the value is, the better is the detection performance of the given algorithm.

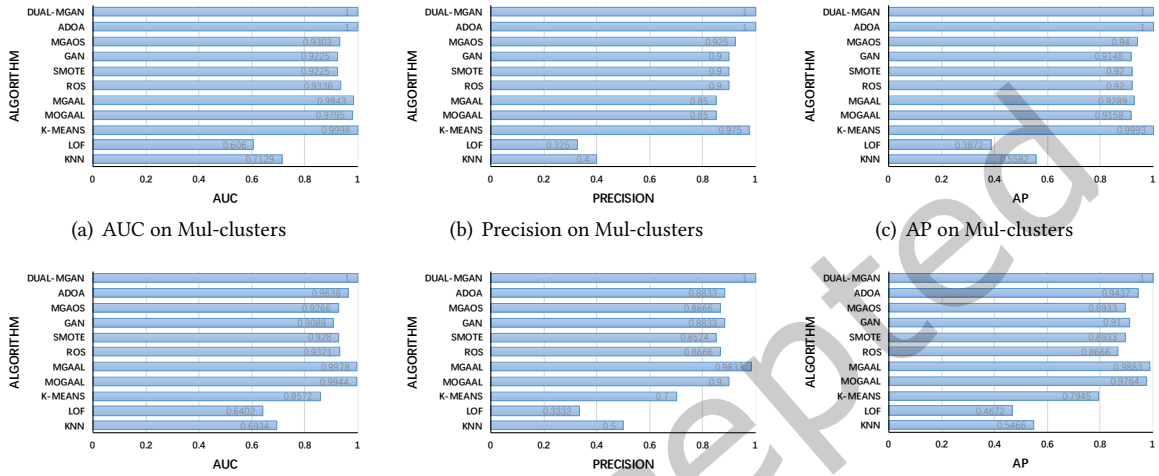
5.1.3 Baselines. To evaluate the detection performance of Dual-MGAN in the case of semi-supervised outlier detection with few identified anomalies and a large amount of unlabeled data, we compared it with eight representative outlier detection algorithms. (i) Three of the most common unsupervised approaches (i.e., kNN, LOF, and k-means) and the original version of MGAAL (i.e., MOGAAL) were used to investigate the significance of the use of identified anomalies. (ii) Three representative oversampling-based supervised methods (i.e., ROS, SMOTE, GAN) were used as baselines to explore the significance of establishing normal patterns. (iii) The semi-supervised ADOA, which attaches a weight to each instance according to its isolation and similarity, was used to further verify the performance advantages of our proposed semi-supervised model. In addition, the two sub-modules (i.e., MGAAL and MGAOS) in Dual-MGAN were separately tested to illustrate their effectiveness as an unsupervised detector and a supervised detector, respectively, as well as to show the necessity of combining these two sub-modules.

5.1.4 Parameter Settings. For the three common unsupervised algorithms, the identified anomalies were removed from the training data to use their explicit information. The optimal parameters were searched for in a range of values, such as the parameters k in the kNN and LOF were searched for from 2 to $\min(\lceil \frac{n}{10} \rceil, 100)$, and k in k-means was searched for in the range of 1 to $\min(\lceil \frac{n}{100} \rceil, 10)$. For supervised anomaly detection, we sampled or generated a large number of samples to convert the original datasets into balanced datasets. An artificial neural network was then introduced to the training process as an outlier detector. For the semi-supervised algorithm, the ratio β in ADOA was changed from 0.1 to 0.9 to obtain the best results. For MOGAAL, Dual-MGAN, and the two proposed sub-modules, we used the following network structure: (i) The output value was used to divide the dataset in MOGAAL, and we set k to 10. (ii) k-means clustering was used to divide the dataset in others, and we set k_a and k_u to $\min(n_a, 10)$ and $\min(n_u, 10)$, respectively. (iii) We used a three-layer network $(d * d * d)$ for the sub-generator, and a four-layer network $(d * \min(n, 1000) * 10 * 1)$ for the sub-discriminator and the detector D. (iv) We used an Orthogonal initializer for the generator, and the Variance-scaling for the sub-discriminator and the detector D. (v) The learning rates of the sub-generator and the sub-discriminator were set to 0.0001 and 0.1, respectively, and that of the detector was tuned in the range of $\{0.01, 0.001, 0.0001\}$. (vi) We directly collected similar samples for subsets containing fewer than ten identified anomalies, and the thresholds of Nnr_{aj} and Nnr_{uj} were set to 0.4 and 0.2, respectively. Moreover, for all algorithms that used artificial neural networks as detectors, the AUC was used to select the final model.

5.2 Experimental Results

5.2.1 Experimental Results on Synthetic Data. The experimental results of Dual-MGAN and the ten competitors are shown in Fig. 9. Dual-MGAN obtained perfect results on both the Mul-clusters and Mul-shapes datasets, that is, the value of all its three indicators was 1. The semi-supervised ADOA and cluster-based k-means obtained

good results on the Mul-clusters (AUC was close to 1), but they did not perform well on the Mul-shapes. The other competitors all obtained mediocre results on the two datasets. In particular, because the kNN and LOF, when the parameter k is in a specific range, cannot identify group anomalies, their detection performance was significantly worse than normal. To more clearly illustrate the characteristics of their performance, we provide a visual representation of the results in Figs. 10 and 11.

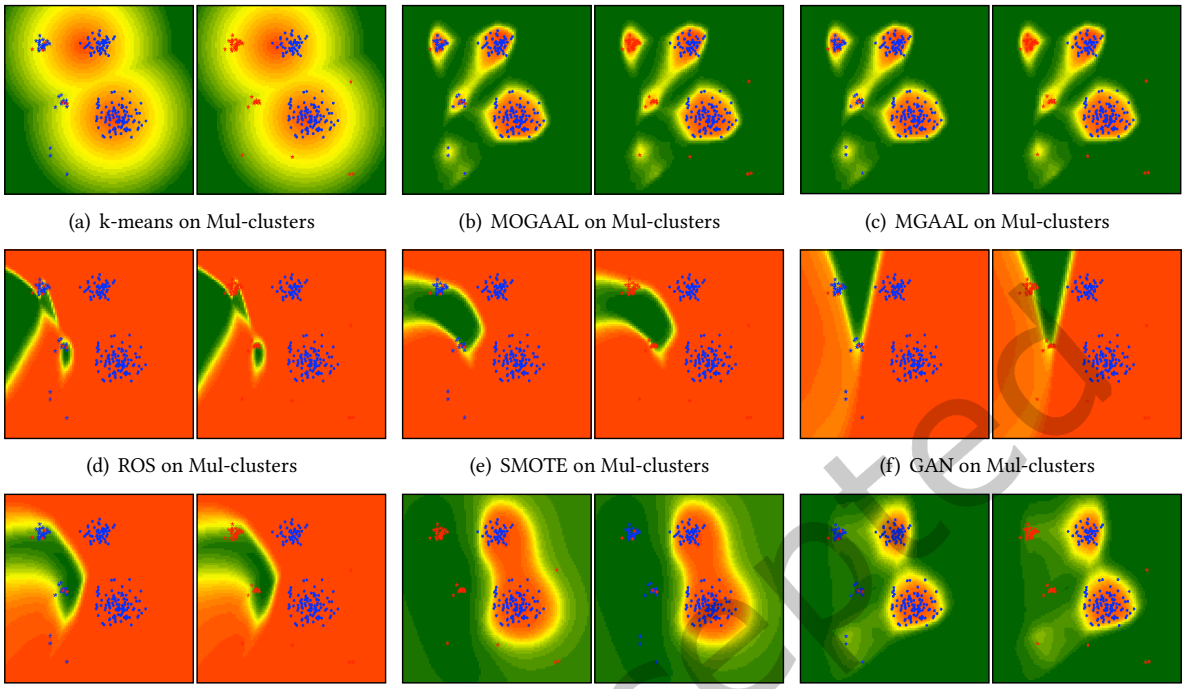


(a) AUC on Mul-clusters (b) Precision on Mul-clusters (c) AP on Mul-clusters

(d) AUC on Mul-shapes (e) Precision on Mul-shapes (f) AP on Mul-shapes

Fig. 9. Experimental results on the two synthetic datasets.

On the Mul-clusters dataset, unsupervised k-means achieved the optimal result when $k = 2$ (as shown in Fig. 10(a)). Although centers of the clusters were not accurately identified due to interference by the unidentified anomalies, partially identified group anomalies and discrete anomalies were all correctly separated from the normal data. But on the Mul-shapes dataset, k-means (as shown in Fig. 11(a)) established an incorrect model because its assumption could not be satisfied by the U-shaped cluster. The two AGPO-based unsupervised approaches (as shown in Figs. 10(b)-10(c) and Figs. 11(b)-11(c)) described a boundary that enclosed the concentrated data such that discrete anomalies could be accurately identified. However, partially identified group anomalies could not be separated from the concentrated normal data because only the explicit information in the identified anomalies was used. Contrary to unsupervised methods, the four supervised algorithms (i.e., ROS, SMOTE, GAN, and MGAOS) made full use of the potential information in the identified anomalies by data augmentation, so that the two sets of group anomalies were effectively identified (shown in Figs. 10(d)-10(g)). However, in addition to partially identified group anomalies, there were unidentified discrete anomalies and emerging anomalies in these scenarios. Since supervised detectors only separated the identified abnormal patterns from other data, instead of establishing normal patterns. They faced substantial challenges in the detection of previously unknown anomalies, which is the weakness compared to the semi-supervised Dual-MGAN. Moreover, compared with the supervised sub-module (i.e., MGAOS) in Dual-MGAN, the over-fitting problem in ROS and the over-generalization problem in SMOTE were shown in Fig. 10(d) and Fig. 11(e), respectively. As for the semi-supervised ADOA, all anomalies in the training and test data of Mul-clusters can be identified (as shown in Fig. 10(h)). However, some emerging anomalies in the test data of Mul-shapes were incorrectly detected as normal data (as shown in Fig. 11(h)). This is because ADOA divided only the weighted normal data from the weighted anomalies, such that the detection results of emerging anomalies in the test data could not be guaranteed. By contrast, our proposed Dual-MGAN (as shown in Fig. 10(i) and Fig. 11(i)) detected all anomalies and described the correct boundary by



(a) k-means on Mul-clusters (b) MOGAAL on Mul-clusters (c) MGAAL on Mul-clusters

(d) ROS on Mul-clusters (e) SMOTE on Mul-clusters (f) GAN on Mul-clusters

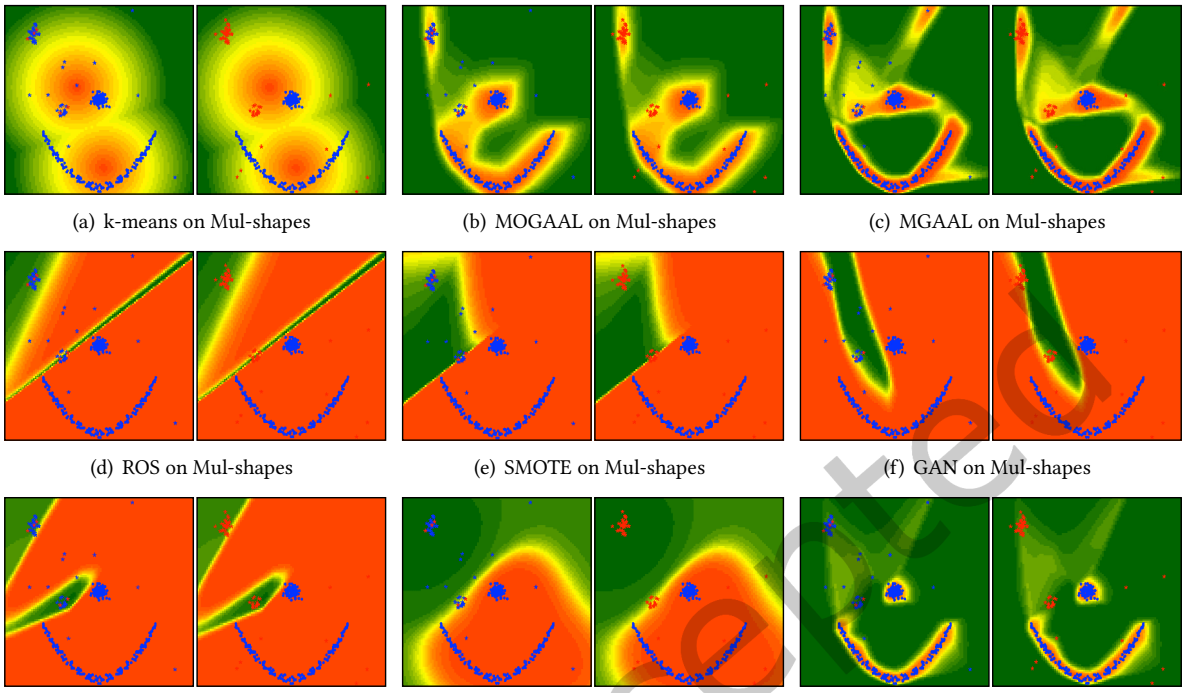
(g) MGAOS on Mul-clusters (h) ADOA on Mul-clusters (i) Dual-MGAN on Mul-clusters

Fig. 10. Detection results of different algorithms on the Mul-clusters dataset.

simultaneously using the potential information and constructing the normal patterns. Moreover, compared with ADOA, Dual-MGAN had a significant advantage in handling data clusters of various shapes.

5.2.2 Experimental Results on Real-world Data. The AUC, precision, and AP of different algorithms on the real-world datasets are shown in Tables 2-4. The best result for each dataset is highlighted in bold, and the average ranks of 11 algorithms on the ten datasets are provided in the last row. In general, the proposed Dual-MGAN obtained the best overall performance (i.e., average rank) on all three evaluation measures. The semi-supervised ADOA could identify partially identified group anomalies and discrete anomalies in the training data. However, owing to the significant challenge in detecting emerging anomalies, its overall performance was inferior to that of the semi-supervised Dual-MGAN. The unsupervised and supervised algorithms had their own advantages and disadvantages on different datasets, and their overall performance was not as good as that of the proposed Dual-MGAN. But it is worth noting that the proposed sub-modules (i.e., MGAAL and MGAOS) obtained better results than similar algorithms.

To further explore the improvement effected by the proposed Dual-MGAN on different datasets, the trends of performance with different ratios of identified anomalies are shown in Fig. 12. Dual-MGAN used few identified anomalies (i.e., 10% identification ratio) to achieve excellent results that approached those obtained when all tags are known (i.e., 100% identification ratio) on multiple datasets. Because the performance of similar algorithms is not exactly the same, we focus on the difference between Dual-MGAN and the two sub-modules. This can also help eliminate the influence of other factors. The difference between MGAAL without any identified anomalies and MGAAL with few identified anomalies shows the improvement brought about by using explicit information and guidance information in the identified anomalies. The difference between MGAAL and Dual-MGAN with the



(a) k-means on Mul-shapes (b) MOGAAL on Mul-shapes (c) MGAAL on Mul-shapes

(d) ROS on Mul-shapes (e) SMOTE on Mul-shapes (f) GAN on Mul-shapes

(g) MGAOS on Mul-shapes (h) ADOA on Mul-shapes (i) Dual-MGAN on Mul-shapes

Fig. 11. Detection results of different algorithms on the Mul-shapes dataset.

Table 2. Detection performance (AUC) of different algorithms on real-world datasets

Dataset	kNN	LOF	k-means	MOGAAL	MGAAL	ROS	SMOTE	GAN	MGAOS	ADOA	Dual-MGAN
Pima	0.7461	0.7380	0.6889	0.7178	0.7474	0.7052	0.6770	0.5173	0.7286	0.6961	**0.7873**
Stamps	0.9339	0.9203	0.9019	0.9456	0.9419	0.8988	0.8972	0.9092	0.9108	0.9213	**0.9964**
Shuttle	0.7906	0.5771	0.9854	0.9880	0.9886	0.9825	0.9808	0.9779	0.9849	**0.9928**	0.9925
Cover	0.9132	0.9115	0.9468	0.9838	0.9838	0.9994	0.9993	0.9764	**0.9995**	0.9980	**0.9995**
Vowels	0.9764	**0.9785**	0.9482	0.9150	0.9672	0.9095	0.9002	0.8661	0.9040	0.9428	0.9517
Cardio	0.9598	0.9280	0.9588	0.8988	0.9629	0.9905	0.9872	0.8735	**0.9936**	0.9853	0.9930
WDBC	0.9186	0.9214	0.9132	0.9475	0.8518	0.9349	0.9132	0.9322	0.9367	0.9214	**0.9520**
Ionosphere	0.9330	0.9215	0.9209	0.9151	0.9450	0.6800	0.6172	0.6595	0.7582	0.8777	**0.9538**
Arrhythmia	0.7290	0.7243	0.7173	0.7457	0.7507	0.6625	0.5736	0.6511	0.6411	0.5965	**0.7585**
Speech	0.5424	0.6416	0.4856	0.5599	0.5645	0.6586	0.6602	0.4448	0.6754	0.6713	**0.6899**
A. R.	6.0	6.3	7.8	5.2	4.5	6.2	8.2	9.2	5.0	5.7	**1.5**

same ratio of identified anomalies shows the improvement effected by data augmentation. The difference between MGAOS and Dual-MGAN with the same ratio of identified anomalies shows the significance of establishing normal patterns.

Table 3. Detection performance (precision) of different algorithms on real-world datasets

Dataset	kNN	LOF	k-means	MOGAAL	MGAAL	ROS	SMOTE	GAN	MGAOS	ADOA	Dual-MGAN
Pima	0.5393	0.5617	0.5280	0.5180	0.5517	0.5317	0.5318	0.3782	0.5692	0.5168	**0.6067**
Stamps	0.4000	0.4000	0.4000	0.5666	0.4666	**0.9000**	0.8000	0.8333	**0.9000**	**0.9000**	**0.9000**
Shuttle	0.1441	0.1144	0.9519	0.9551	0.9527	0.958	**0.9571**	0.9467	0.9583	0.9467	0.9542
Cover	0.1035	0.1542	0.1409	0.3468	0.5016	0.8703	0.8678	0.5329	0.8692	0.7235	**0.8839**
Vowels	0.4375	0.4375	0.3125	0.4583	0.6041	0.5833	0.5208	0.3333	0.5625	0.4375	**0.6666**
Cardio	0.6206	0.4827	0.6724	0.6953	0.7183	0.8562	0.8333	0.5172	**0.8792**	0.7931	0.8563
WDBC	0.3333	**0.6666**	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.4444	0.3333	**0.6666**
Ionosphere	0.8095	0.7619	**0.8571**	0.8095	0.8333	0.3589	0.3756	0.3979	0.6169	0.7619	**0.8571**
Arrhythmia	0.6470	0.6323	0.6176	**0.6862**	0.6813	0.6127	0.5244	0.5539	0.5784	0.5441	0.6715
Speech	0.0500	0.0500	0.0000	0.1000	0.1160	0.1160	0.1333	0.0000	0.1166	**0.1500**	0.1333
A. R.	7.1	7.0	7.4	5.8	4.7	4.5	5.4	8.2	3.6	5.6	**1.8**

Table 4. Detection performance (AP) of different algorithms on real-world datasets

Dataset	kNN	LOF	k-means	MOGAAL	MGAAL	ROS	SMOTE	GAN	MGAOS	ADOA	Dual-MGAN
Pima	0.5732	0.5545	0.5391	0.5431	0.5858	0.5266	0.5546	0.3896	0.5587	0.5188	**0.6219**
Stamps	0.5094	0.4525	0.4212	0.6064	0.5358	0.8914	0.8767	0.8082	0.9017	0.9011	**0.9636**
Shuttle	0.2110	0.0984	0.9551	0.9663	0.9588	0.9711	0.9690	0.9566	0.9739	**0.9755**	0.9746
Cover	0.0718	0.1005	0.1198	0.2854	0.3712	0.9548	0.9483	0.5593	0.9571	0.8132	**0.9612**
Vowels	0.4430	0.5488	0.3363	0.4868	0.4844	0.5874	0.5721	0.3972	0.6108	0.5198	**0.7190**
Cardio	0.6791	0.4730	0.6872	0.7047	0.7641	0.9372	0.9225	0.5485	**0.9514**	0.8890	0.9361
WDBC	0.3666	0.5878	0.4619	0.5088	0.1661	0.3928	0.3323	0.5384	0.4695	0.5333	**0.6445**
Ionosphere	0.9166	0.8744	0.9253	0.9066	0.9200	0.6086	0.5938	0.6098	0.7329	0.8279	**0.9310**
Arrhythmia	0.7449	0.7414	0.7329	0.7494	0.7593	0.6640	0.5696	0.6516	0.6543	0.5797	**0.7621**
Speech	0.0232	0.0389	0.0189	0.0416	0.0566	**0.1336**	0.1186	0.0155	0.1127	0.1331	0.1243
A. R.	7.8	7.5	8.2	6.3	6.1	5.2	6.4	7.4	4.2	5.4	**1.5**

Specifically, on the Cover and Cardio datasets, the supervised MGAOS obtained the same results as the semi-supervised Dual-MGAN by making full use of the potential information in the identified anomalies (as shown in Table 2). This is because the identified anomalies in these datasets represented almost all anomalous patterns. In this case, the improvement effected by data augmentation was obviously, and the unsupervised MGAAL obtained similar results only when there were a large number of identified anomalies in these datasets (i.e., more than 50% identification ratio, as shown in Figs. 12(d) and 12(f)). However, on some other datasets, the detection performance of supervised algorithms was significantly worse than that of unsupervised and semi-supervised algorithms. This is because the anomalies in these datasets were relatively discrete, and anomalous patterns could not be completely represented by the identified anomalies. In particular, when all identified anomalies were discrete, data augmentation led to almost no improvement in performance, as evidenced by the results on the Vowels, Ionosphere, and Arrhythmia datasets (as shown in Table 2). In this case, it was of great significance to establish normal patterns, and the supervised MGAOS obtained similar results only when all anomalies were identified (i.e., 100% identification ratio, as shown in Figs. 12(e), 12(h), and 12(i)). However, in addition to the two extreme cases

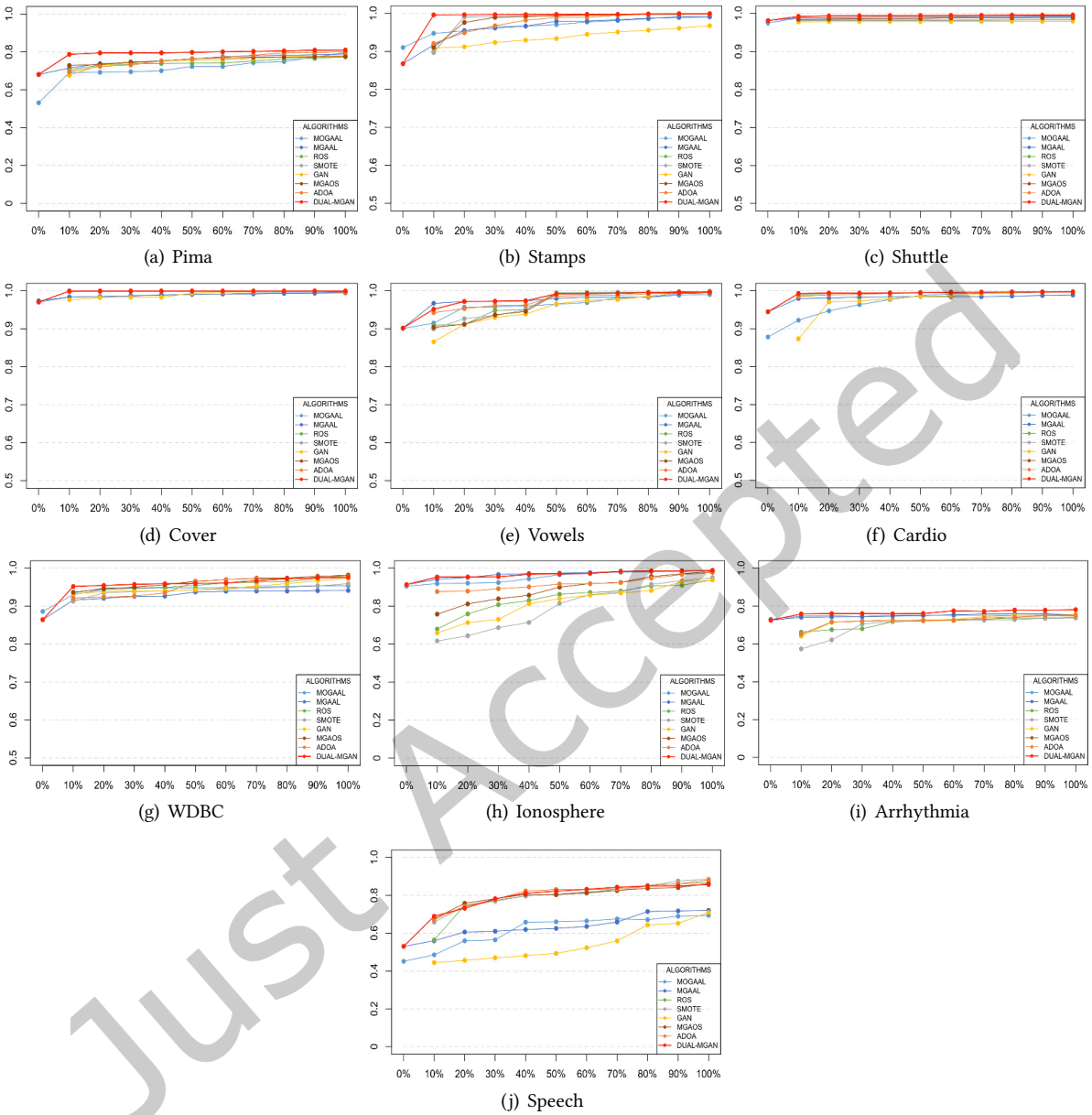


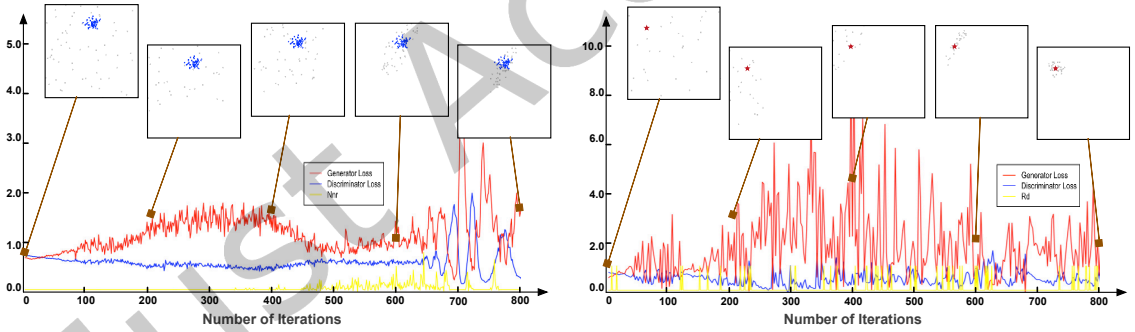
Fig. 12. Performance trends of different algorithms on real-world datasets with different identification ratios. The ratio of identified anomalies in each dataset was adjusted from 0% to 100%.

mentioned above, the dataset usually contains both discrete anomalies and group anomalies, such as the synthetic datasets considered in Section 5.1.1. In this case, the semi-supervised Dual-MGAN obtained a higher precision by combining the two sub-modules, which were respectively used to detect discrete anomalies and partially identified group anomalies. Both the supervised and the unsupervised algorithms required significantly more

identified anomalies to detect the corresponding anomalous patterns (as shown in Figs. 12(a)-12(c) and 12(g)). In general, the proposed Dual-MGAN achieved excellent results by simultaneously using the potential information and constructing the normal patterns. Although the supervised and unsupervised algorithms obtained satisfactory results in some extreme cases, the semi-supervised Dual-MGAN was more robust for various data distributions.

5.3 Evaluation Indicators

Because all sub-GANs in Dual-MGAN stop training once they can provide a reasonable reference distribution or augment the minority class, the evaluation of the training status has a substantial effect on the results. Therefore, additional experiments were conducted on both synthetic and real-world datasets to investigate the effectiveness of the two evaluation indicators: the nearest neighbor ratio Nnr and relative distance Rd. The optimization process of two sub-GANs on the Mul-clusters dataset is shown in Fig. 13. The GAN_{ui} in MGAAL gradually learned the generation mechanism of the unlabeled data (shown with blue dots in Fig. 13(a)), and the generated potential outliers (shown with gray dots in Fig. 13(a)) were aggregated in the area where the unlabeled data were located after 400 iterations. Meanwhile, the value of Nnr was almost zero in the first 400 iterations, and then increased significantly in the subsequent 200 iterations (shown with the yellow line in Fig. 13(a)). This shows that the value of Nnr can reflect the training status of GAN_{ui}, and the generated potential outliers provide a reasonable reference distribution when the value of the indicator is greater than the threshold. For the subset containing only one data point (shown with red star in Fig. 13(b)), the value of Rd was sometimes 1 due to the randomness of a single generated data item. However, the frequency of $Rd = 1$ increased as training progressed, and the potential outliers generated by GAN_{ai} prevented the data point from being ignored when Rd was frequently equal to 1. This shows that the value of Rd can reflect the training status of GAN_{ai} when the subset contains only one data point. Moreover, to better guarantee the performance of Dual-MGAN, we directly collected a certain number of generated data items whose Rd was 1 (shown with gray dots in Fig. 13(b)).

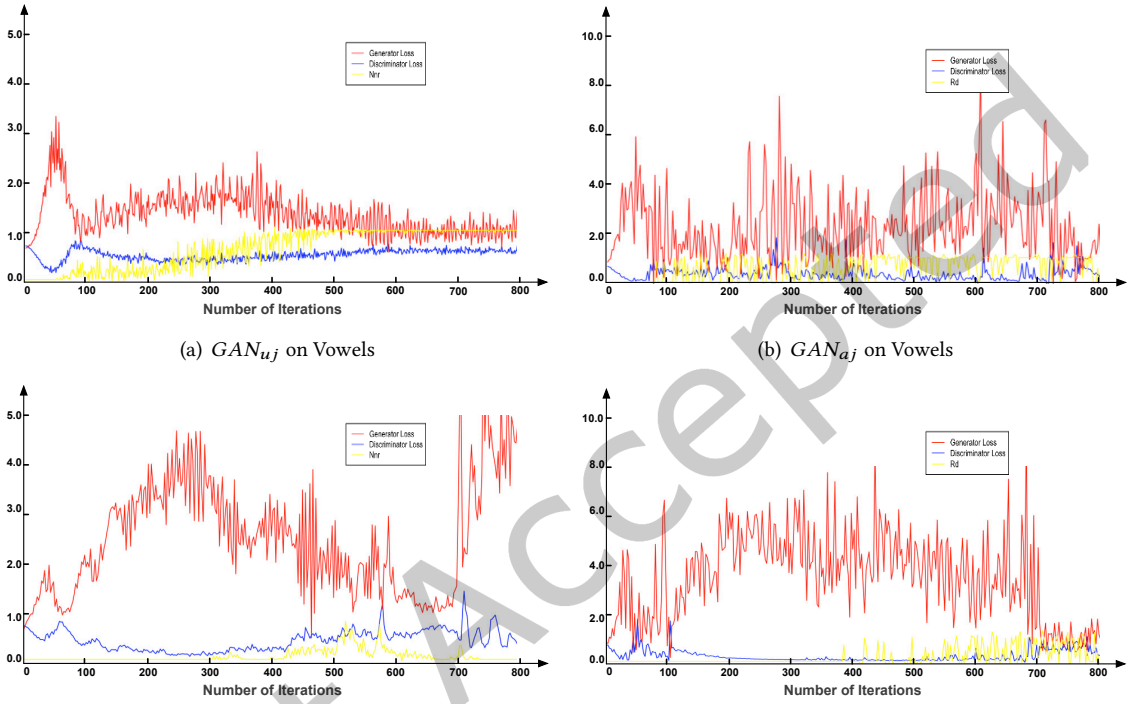


(a) GAN_{ui} on Mul-clusters (b) GAN_{ai} on Mul-clusters

Fig. 13. The optimization process of sub-GANs on the Mul-clusters dataset. The loss of the sub-generator is shown using the red line, that of the sub-discriminator using the blue line, and the value of the indicator is represented by the yellow line.

In addition to the two evaluation indicators, the sub-generator loss can also reflect the training status of the sub-GANs (shown with the red line in Fig. 13). That is, the sub-GAN can generate informative potential outliers when the downward trend of the sub-generator loss tends to be slow. Because it is difficult to visualize high-dimensional data, the sub-generator loss was used to verify the effectiveness of the proposed evaluation indicators on the real-world datasets. Fig. 14 displays the optimization process of sub-GANs on the Vowel and Arrhythmia datasets, where the sub-generator loss showed a relatively prominent trend. On the Vowel dataset, the value of Nnr was greater than the threshold when the sub-generator loss stopped decreasing (as shown in Fig.

14(a)), and the frequency of $Rd = 1$ increased when the downward trend of the sub-generator loss tended to be slow (as shown in Fig. 14(b)). On the Arrhythmia dataset, both evaluation indicators reached the stop conditions when the downward trends of sub-generator loss tended to be slow after the second peak (as shown in Figs. 14(c) and 14(d)). The consistency with the trend of loss shows the effectiveness of the evaluation indicators on high-dimensional data. Moreover, compared with the generator loss, Nnr can more easily adjust the final state of the sub-GANs by setting different thresholds without requiring human intervention.



(a) GAN_{uj} on Vowels

(b) GAN_{aj} on Vowels

(c) GAN_{uj} on Arrhythmia

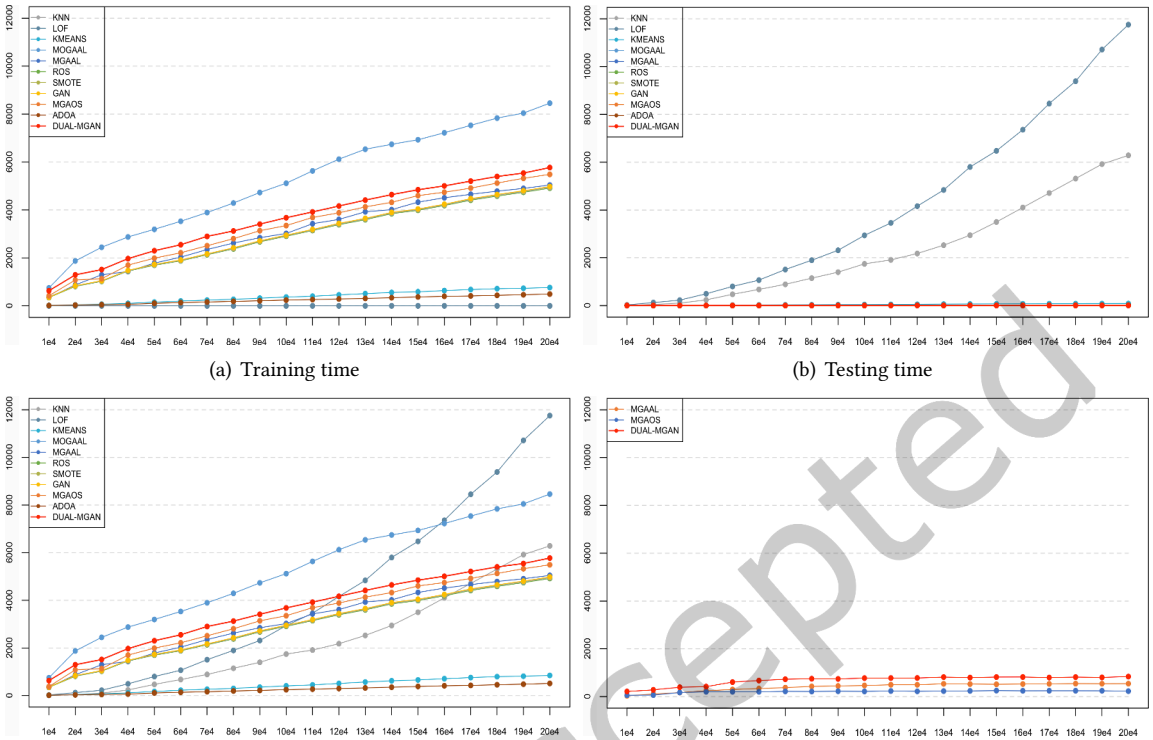
(d) GAN_{aj} on Arrhythmia

Fig. 14. The optimization process of sub-GANs on the two real-world datasets.

5.4 Computational Complexity

Although neural network models generally have high computational complexity, they are being applied in a growing number of fields owing to their powerful representational capabilities. In this section, the computational complexity of different algorithms is discussed so that researchers can make informed choices of models based on specific situations. The training and testing times of different algorithms on the Cover dataset are shown in Fig. 15, where the horizontal axis represents the amount of data extracted from the dataset, and the vertical axis represents the run time.

Unsupervised kNN and LOF did not require training, but it was necessary to measure the distance or density ratio between all test data and the training data during the test. Therefore, their testing times (shown with gray and navy lines in Fig. 15(b), respectively) increased significantly with the amount of data, which is not suitable for real-time online detection. The clustering-based k-means established normal patterns through a heuristic algorithm, such that its training and testing times were short. Therefore, when the number of clusters is known



(a) Training time

(b) Testing time

(c) Total time

(d) Training time of sub-GANs

Fig. 15. The run times of different algorithms on the Cover dataset.

and all clusters follow a Gaussian distribution, k-means is a good choice. The training times of AGPO-based unsupervised algorithms (i.e., MOGAAL and MGAAL) were significantly higher than those of k-means, but they could handle data clusters of various shapes without knowing the number of clusters. Therefore, when there are no group anomalies, and the clusters do not all follow a Gaussian distribution, MOGAAL and MGAAL are a good choice. Furthermore, as MOGAAL must calculate the target value in each iteration, the calculation cost of MGAAL was lower than that of MOGAAL.

There was no significant difference in terms of run time among the four supervised algorithms. Although the over-sampling in ROS and SMOTE took little time, they all introduced an artificial neural network to the training process as an outlier detector. Each iteration of the neural network consisted of parameter learning and model evaluation. As the neural networks were trained using mini-batches, the computational cost of parameter learning was low, and barely increased with the amount of data. But the model evaluation required calculating the AUC of the output such that the training time increased linearly with n. The difference between MGAOS and ROS was the computational cost of using multiple sub-GANs for over-sampling (shown with the blue line in Fig. 15(d)). The run time of over-sampling in MGAOS increased linearly with n_a, and barely changed when the number of identified abnormalities reached a certain number. This is because the number of iterations was constant and the sub-GANs were trained using mini-batches. From this perspective, the computational complexity of MGAOS-based over-sampling was not too high, especially on large datasets. Therefore, when the identified anomalies can represent all abnormal patterns, and there are no emerging anomalies, MGAAL is a good choice.

However, when the data contain both discrete and group anomalies, or the distribution of the data is unknown, outlier detection using only a supervised or an unsupervised model may not yield satisfactory results.

The semi-supervised ADOA had significant computational advantages over the other methods on all datasets because it was composed of four traditional machine learning modules. Therefore, when the test data do not contain emerging anomalies and the distribution of data is not complex, ADOA is a wonderful choice. The semi-supervised Dual-MGAN had a slightly higher computational cost than the supervised algorithms, which was significantly higher than that of ADOA. But it could identify a correct boundary even in cases of emerging anomalies and complicated data distribution (as shown in Fig. 11(i)). In addition, the difference between Dual-MGAN and ROS was the computational cost of using multiple sub-GANs to construct the distribution and augment the data (shown with the red line in Fig. 15(d)). The cost was low and barely changed when n reached a certain value. And because the performance of the detector on a subset can represent its performance on the full set in many cases, there is considerable room to optimize the computational cost of the evaluation. Therefore, the high computational complexity of Dual-MGAN is not absolute, as are other algorithms that use artificial neural network as a detector. When the distribution of the data is complex or unknown, Dual-MGAN is a reliable choice.

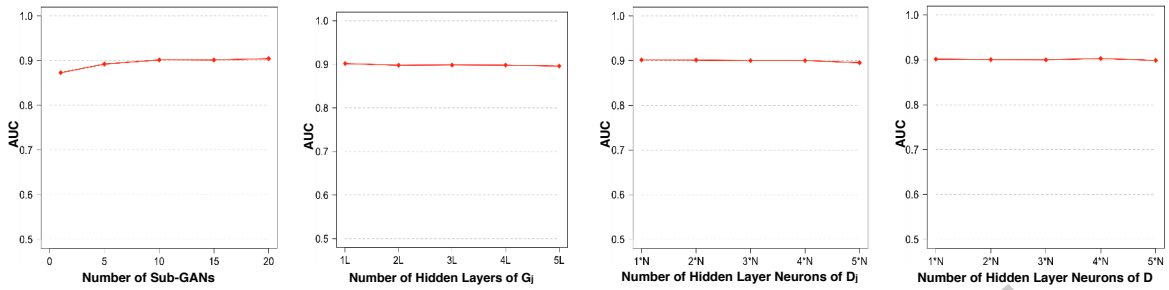
5.5 Robustness Analysis

Dual-MGAN involves several parameters, as described in Section 5.1.4. Some parameters (e.g., initializer, learning rates, and thresholds) are carefully set for the detection task, while the parameter setting related to the network structure is relatively uncomplicated. Therefore, this section focuses on the influence of network structure-related parameters on the detection results, mainly including the number of sub-GANs, the number of hidden layers of the sub-generator, the number of hidden layer neurons of the sub-discriminator, and the number of hidden layer neurons of the detector. In addition, since the precision and average precision fluctuate greatly when the dataset contains only a few anomalies, the average values of AUC obtained by different network structures on all real-world datasets are used to analyze the robustness of these parameters.

The numbers of sub-GANs (i.e., k_a and k_u) determine the level of detail of distribution construction and data augmentation, respectively. To evaluate their effect on performance, k_a and k_u were set in the range of {1, 5, 10, 15, 20}, and the results of Dual-MGAN using different parameters on the real-world datasets are shown in Fig.16(a). Due to the mode collapse problem, the overall performance of Dual-MGAN with only one sub-GAN (i.e., $k_a = k_u = 1$) was lower than that of other models with multiple sub-GANs. However, as the numbers of sub-GANs increased, the performance of the model increased and eventually remained stable. This implies that the number of sub-GANs is robust when reaching a certain size. As for the internal structure of each neural network, the number of hidden layers of the sub-generator determines the learning ability of G_{aj}/G_{uj}, and the numbers of hidden layer neurons of the sub-discriminator and detector determine the representation power of the D_{aj}/D_{uj} and D. To evaluate their robustness, we adjusted the number of hidden layers (from 1 to 5) and that of hidden layer neurons (from $\min(n, 1000)$ to $5 * \min(n, 1000)$), respectively. The results of Dual-MGAN with different network structures on the real-world datasets are shown in Fig. 16(b)-16(d). Although there might be differences in the performance of different models on individual datasets, their overall performance was almost the same. This shows that the proposed Dual-MGAN is not sensitive to the above parameters, and the parameters set in Section 5.1.4 is able to obtain acceptable performance.

6 CONCLUSIONS AND FUTURE WORKS

This article focused on semi-supervised outlier detection with the aim of using few identified anomalies and a large amount of unlabeled data to identify as many anomalies as possible. This special task of outlier detection was first decomposed into the detection of discrete anomalies and that of partially identified group anomalies. For the first sub-task, the distribution construction sub-module (i.e., MGAAL) was proposed to construct a reasonable



(a) (b) (c) (d)

Fig. 16. Experimental results of different network structures on real-world datasets. Vertical axis represents the average result of a particular network structure on all above real-world datasets.

reference distribution to ensure that discrete anomalies could be separated from concentrated normal data. For the second sub-task, a data augmentation sub-module (i.e., MGAOS) was proposed to increase the size of the minority class to ensure that both the identified anomalies and the partially identified group anomalies were detected. In this way, the semi-supervised Dual-MGAN that combines MGAAL and MGAOS can not only identify discrete anomalies, but can also detect partially identified group anomalies accurately. Moreover, to obtain more reliable results, the AUC of the output was used to select the final model, and two evaluation indicators based on relative distance were introduced to evaluate the training status of the sub-GANs. The results of extensive experiments on synthetic data and real-world data showed that our proposed approach can usually improve the accuracy of outlier detection by simultaneously using the potential information and constructing the normal patterns. Although other algorithms obtained satisfactory results in some extreme cases, the semi-supervised Dual-MGAN was more robust for various data distributions. Additional experiments on the evaluation indicators showed that the proposed *Nnr* and *Rd* can reflect the training status of the sub-GANs. In the context of computational complexity, because there is still considerable room for optimizing its computational cost, the high computational complexity of Dual-MGAN is not absolute.

The proposed model can be regarded as a general framework for all semi-supervised outlier detection with few identified anomalies. However, the network structure of the sub-generator in this article cannot generate unstructured data, such as text, voice, and images. Therefore, a deeper study on the network structures for different data types will be explored for a wider application area. Another interesting direction for future work is interpretability. In addition to the abnormal value of the data, the reason why an instance is considered anomalous is important for most applications. Therefore, in order to provide users with more valuable information, more intensive research will be conducted on the interpretation of anomaly detectors.

REFERENCES

[1] C. C. Aggarwal. 2017. *Outlier Analysis*. Springer International Publishing.

[2] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon. 2018. GANomaly: Semi-supervised Anomaly Detection via Adversarial Training. In *Computer Vision – ACCV 2018*. 622–637. https://doi.org/10.1007/978-3-030-20893-6_39

[3] F. Angiulli. 2019. CFOF: A Concentration Free Measure for Anomaly Detection. *ACM Transactions on Knowledge Discovery from Data* 14, 1 (Feb. 2019), 1–53. https://doi.org/10.1145/3362158

[4] A. Belhadi, Y. Djenouri, and J. C. Lin. 2019. Comparative Study on Trajectory Outlier Detection Algorithms. In *International Conference on Data Mining Workshops (ICDMW)*. Beijing China, 415–423. https://doi.org/10.1109/ICDMW.2019.00067

[5] J. Bian, X. L. Hui, S. Y. Sun, X. G. Zhao, and M. Tan. 2019. A Novel and Efficient CVAE-GAN-Based Approach with Informative Manifold for Semi-Supervised Anomaly Detection. *IEEE Access* 7 (2019), 88903–88916. https://doi.org/10.1109/ACCESS.2019.2920251

[6] G. O. Campos, A. Zimek, J. Sander, R. J. Campello, M. Barbora, E. Schubert, I. Assent, and M. E. Houle. 2016. On The Evaluation of Unsupervised Outlier Detection: Measures, Datasets, and An Empirical study. *Data Mining and Knowledge Discovery* 30, 4 (2016), 891–927. https://doi.org/10.1007/s10618-015-0444-8

[7] M. H. Chehreghani. 2016. K-Nearest Neighbor Search and Outlier Detection via Minimax Distances. In *SIAM International Conference on Data Mining*. 405–413. https://doi.org/10.1137/1.9781611974348.46
[8] D. W. Cheng, X. Y. Wang , Y. Zhang, and L. Q. Zhang. 2020. Graph Neural Network for Fraud Detection via Spatial-temporal Attention. *IEEE Transactions on Knowledge and Data Engineering* (2020), 1–1. https://doi.org/10.1109/TKDE.2020.3025588
[9] A. Daneshpazhouh and A. Sami. 2013. Semi-supervised Outlier Detection with Only Positive and Unlabeled Data Based on Fuzzy Clustering. In *The 5th Conference on Information and Knowledge Technology*. Shiraz Iran, 344–348. https://doi.org/10.1109/IKT.2013.6620091
[10] A. Daneshpazhouh and A. Sami. 2014. Entropy-Based Outlier Detection Using Semi-supervised Approach with Few Positive Examples. *Pattern Recognition Letters* 49 (2014), 77–84. https://doi.org/10.1016/j.patrec.2014.06.012
[11] K. Ghosh Dastidar, J. Jurgovsky, W. Siblini, L. He-Guelton, and M. Granitzer. 2020. NAG: Neural Feature Aggregation Framework for Credit Card Fraud Detection. In *2020 IEEE International Conference on Data Mining (ICDM)*. Sorrento Italy, 92–101. https://doi.org/10.1109/ICDM50108.2020.00018
[12] Y. Dou, G. Ma, P. S. Yu, and S. Xie. 2020. Robust Spammer Detection by Nash Reinforcement Learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 924–933. https://doi.org/10.1145/3394486.3403135
[13] S. M. Erfani, S. Rajasegarar, S. Karunasekera, and C. Leckie. 2016. High-Dimensional and Large-Scale Anomaly Detection Using a Linear One-Class SVM with Deep Learning. *Pattern Recognition* 58 (Oct. 2016), 121–134. https://doi.org/10.1016/j.patcog.2016.03.028
[14] T. Ergen and S. S. Kozat. 2020. Unsupervised Anomaly Detection with LSTM Neural Networks. *IEEE Transactions on Neural Networks and Learning Systems* 31, 8 (Aug. 2020), 3127–3141. https://doi.org/10.1109/TNNLS.2019.2935975
[15] U. Fiorea, A. D. Santis, F. Perla, P. Zanetti, and F. Palmieri. 2017. Using Generative Adversarial Networks for Improving Classification Effectiveness in Credit Card Fraud Detection. *Information Sciences* 479 (April 2017), 448–455. https://doi.org/10.1016/j.ins.2017.12.030
[16] J. Gao, H. B. Cheng, and P. N.Tan. 2006. Semi-supervised Outlier Detection. In *ACM symposium on Applied computing*. 635–636. https://doi.org/10.1145/1141277.1141421
[17] Y. D. Gao, B. Shi, B. Dong, Y. Y. Wang, L. Y. Mi, and Q. H. Zheng. 2021. Tax Evasion Detection with FBNE-PU Algorithm Based on PnCGCN and PU Learning. *IEEE Transactions on Knowledge and Data Engineering* (2021), 1–1. https://doi.org/10.1109/TKDE.2021.3090075
[18] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative Adversarial Networks. *Advances in Neural Information Processing Systems* 3 (2014), 2672–2680. https://doi.org/10.1145/3422622
[19] S. Kim, Y. C. Tsai, K. Singh, Y. Choi, and M. Cha. 2020. DATE: Dual Attentive Tree-aware Embedding for Customs Fraud Detection. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2880—-2890. https://doi.org/10.1145/3394486.3403339
[20] M. Kimura and T. Yanagihara. 2018. Anomaly Detection Using GANs for Visual Inspection in Noisy Training Data. In *Computer Vision – ACCV 2018 Workshops*. Springer Cham, 373–385. https://doi.org/10.1007/978-3-030-21074-8_31
[21] Y. Li, P. Hu, J. Z. Liu, D. Peng, J. T. Zhou, and X. Peng. 2020. Contrastive Clustering. *CoRR* abs/2009.09687 (2020). https://arxiv.org/abs/2009.09687
[22] H. J. Liao, C. Lin, Y. C. Lin, and K. Y. Tung. 2013. Intrusion detection system: A comprehensive review. *Journal of Network and Computer Applications* 36, 1 (2013), 16–24. https://doi.org/10.1016/j.jnca.2012.09.004
[23] Q. Liao, H. Y. Chai, H. Han, X. Zhang, X. Wang, W. Xia, and Y. Ding. 2021. An Integrated Multi-Task Model for Fake News Detection. *IEEE Transactions on Knowledge and Data Engineering* (2021), 1–1. https://doi.org/10.1109/TKDE.2021.3054993
[24] S. K. Lim, Y. Loo, N. T. Tran, N. M. Cheung, G. Roig, and Y. Elovici. 2018. DOPING: Generative Data Augmentation for Unsupervised Anomaly Detection with GAN. In *IEEE International Conference on Data Mining (ICDM)*. 1122–1127. https://doi.org/10.1109/ICDM.2018.00146
[25] J. L. P. Lima, D. Macêdo, and C. Zanchettin. 2019. Heartbeat Anomaly Detection using Adversarial Oversampling. In *International Joint Conference on Neural Networks (IJCNN)*. 1–7. https://doi.org/10.1109/IJCNN.2019.8852242
[26] R. F. Lima and A. C. M. Pereira. 2017. Feature Selection Approaches to Fraud Detection in e-Payment Systems. In *International Conference on Electronic Commerce and Web Technologies*. 111–126. https://doi.org/10.1007/978-3-319-53676-7_9
[27] B. Liu, Y. S. Xiao, L. B. Cao, Z. F. Hao, and F. Q. Deng. 2013. SVDD-Based Outlier Detection on Uncertain Data. *Knowledge and Information Systems* 34, 3 (March 2013), 597–618. https://doi.org/10.1007/s10115-012-0484-y
[28] B. Liu, Y. S. Xiao, P. S. Yu, Z. F. Hao, and L. B. Cao. 2014. An Efficient Approach for Outlier Detection with Imperfect Data Labels. *IEEE Transactions on Knowledge and Data Engineering* 26, 7 (2014), 1602–1616. https://doi.org/10.1109/TKDE.2013.108
[29] F. T. Liu, K. M. Ting, and Z. H. Zhou. 2008. Isolation Forest. In *IEEE International Conference on Data Mining*. 413–422. https://doi.org/10.1109/ICDM.2008.17
[30] S. H. Liu, B. Hooi, and C. Faloutsos. 2019. A Contrast Metric for Fraud Detection in Rich Graphs. *IEEE Transactions on Knowledge and Data Engineering* 31, 12 (Dec. 2019), 2235–2248. https://doi.org/10.1109/TKDE.2018.2876531
[31] Y. Z. Liu, Z. Li, C. Zhou, Y. C. Jiang, J. S. Sun, M. Wang, and X. N. He. 2020. Generative Adversarial Active Learning for Unsupervised Outlier Detection. *IEEE Transactions on Knowledge and Data Engineering* 32, 8 (2020), 1517–1528. https://doi.org/10.1109/TKDE.2019.2905606
[32] F. Lüer, D. Mautz, and C. Böhm. 2019. Anomaly Detection in Time Series using Generative Adversarial Networks. In *International Conference on Data Mining Workshops (ICDMW)*. 1047–1048. https://doi.org/10.1109/ICDMW.2019.00152

[33] E. Manzoor, S. M. Milajerdi, and L. Akoglu. 2016. Fast Memory-Efficient Anomaly Detection in Streaming Heterogeneous Graphs. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1035–1044. https://doi.org/10.1145/2939672.2939783

[34] J. L. Mao, T. Wang, C. Q. Jin, and A. Y. Zhou. 2017. Feature Grouping-Based Outlier Detection Upon Streaming Trajectories. *IEEE Transactions on Knowledge and Data Engineering* 29, 12 (Dec. 2017), 2696–2709. https://doi.org/10.1109/TKDE.2017.2744619

[35] P. Mishra, V. Varadharajan, U. Tupakula, and E. S. Pilli. 2019. A Detailed Investigation and Analysis of Using Machine Learning Techniques for Intrusion Detection. *IEEE Communications Surveys Tutorials* 21, 1 (2019), 686–728. https://doi.org/10.1109/COMST.2018.2847722

[36] W. A. Mohotti and R. Nayak. 2020. Efficient Outlier Detection in Text Corpus Using Rare Frequency and Ranking. *ACM Transactions on Knowledge Discovery from Data* 14, 6 (Oct. 2020), 1–30. https://doi.org/10.1145/3399712

[37] M. S. Munia, M. Nourani, and S. Houari. 2020. Biosignal Oversampling Using Wasserstein Generative Adversarial Network. In *IEEE International Conference on Healthcare Informatics (ICHI)*. Oldenburg Germany, 1–7. https://doi.org/10.1109/ICHI48887.2020.9374315

[38] M. Odiathevar, W. K.G. Seah, M. Frean, and A. Valera. 2021. An Online Offline Framework for Anomaly Scoring and Detecting New Traffic in Network Streams. *IEEE Transactions on Knowledge and Data Engineering* (2021), 1–1. https://doi.org/10.1109/TKDE.2021.3050400

[39] A. D. Pozzolo, G. Boracchi, O. Caelen, C. Alippi, and G. Bontempi. 2018. Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. *IEEE Transactions on Neural Networks and Learning Systems* 29, 8 (Aug. 2018), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643

[40] P. Qi, J. Cao, T. Y. Yang, J. B. Guo, and J. T. Li. 2019. Exploiting Multi-domain Visual Information for Fake News Detection. In *2019 IEEE International Conference on Data Mining (ICDM)*. Beijing China, 518–527. https://doi.org/10.1109/ICDM.2019.00062

[41] T. Qiu, X. Z. Liu, X. B. Zhou, W. Y. Qu, Z. L. Ning, and C. L. P. Chen. 2020. An Adaptive Social Spammer Detection Model with Semi-supervised Broad Learning. *IEEE Transactions on Knowledge and Data Engineering* (2020), 1–1. https://doi.org/10.1109/TKDE.2020.3047857

[42] Y. X. Ren, B. Wang, J. W. Zhang, and Y. Chang. 2020. Adversarial Active Learning Based Heterogeneous Graph Neural Network for Fake News Detection. In *2020 IEEE International Conference on Data Mining (ICDM)*. Sorrento Italy, 452–461. https://doi.org/10.1109/ICDM50108.2020.00054

[43] M. Sabokrou, M. Khalooei, M. Fathy, and E. Adeli. 2018. Adversarially Learned One-Class Classifier for Novelty Detection. In *IEEE Conference on Computer Vision and Pattern Recognition*. 3379–3388. https://doi.org/10.1109/CVPR.2018.00356

[44] M. Salehi, C. Leckie, J. C. Bezdek, T. Vaithianathan, and X. Y. Zhang. 2016. Fast Memory Efficient Local Outlier Detection in Data Streams. *IEEE Transactions on Knowledge and Data Engineering* 28, 12 (Dec. 2016), 3246–3260. https://doi.org/10.1109/TKDE.2016.2597833

[45] T. Schlegl, P. Seebck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. 2017. Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery. In *International Conference on Information Processing in Medical Imaging*. 146–157. https://doi.org/10.1007/978-3-319-59050-9_12

[46] M. J. Siers and M. Z. Islam. 2021. Class Imbalance and Cost-Sensitive Decision Trees: A Unified Survey Based on a Core Similarity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 15, 1 (Jan. 2021), 1–31. https://doi.org/10.1145/3415156

[47] H. Y. Song, P. Z. Li, and H. F. Liu. 2021. Deep Clustering-based Fair Outlier Detection. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[48] I. Steinwart. 2005. A Classification Framework for Anomaly Detection. *Journal of Machine Learning Research* 6, 1 (2005), 211–232.

[49] B. X. Wang and N. Japkowicz. 2010. Boosting Support Vector Machines for Imbalanced Data Sets. *Knowledge and Information Systems* 25, 1 (2010), 1–20. https://doi.org/10.1007/s10115-009-0198-y

[50] D. X. Wang, J. B. Lin, P. Cui, Q. H. Jia, Z. Wang, Y. M. Fang, Q. Yu, J. Zhou, S. Yang, and Y. Qi. 2019. A Semi-supervised Graph Attentive Network for Financial Fraud Detection. In *IEEE International Conference on Data Mining (ICDM)*. Beijing China, 598–607. https://doi.org/10.1109/ICDM.2019.00070

[51] Y. X. Xie, M. Qiu, H. B. Zhang, L. Z. Peng, and Z. X. Chen. 2020. Gaussian Distribution based Oversampling for Imbalanced Data Classification. *IEEE Transactions on Knowledge and Data Engineering* (April 2020), 1–1. https://doi.org/10.1109/TKDE.2020.2985965

[52] Z. X. Xue, Y. L. Shang, and A. Feng. 2010. Semi-supervised Outlier Detection Based on Fuzzy Rough C-means Clustering. *Knowledge and Information Systems* 80, 9 (May 2010), 1911–1921. https://doi.org/10.1016/j.matcom.2010.02.007

[53] X. Yang, L. J. Latecki, and D. Pokrajac. 2009. Outlier Detection with Globally Optimal Exemplar-Based GMM. In *SIAM International Conference on Data Mining*. 145–154. https://doi.org/10.1137/1.9781611972795.13

[54] X. W. Yi, X. D. Yang, Y. Y. Huang, S. Y. Ke, J. B. Zhang, T. R. Li, and Y. Zheng. 2021. Gas-Theft Suspect Detection among Boiler Room Users: A Data-Driven Approach. *IEEE Transactions on Knowledge and Data Engineering* (2021), 1–1. https://doi.org/10.1109/TKDE.2021.3062707

[55] W. Yu, C. Wei, C. C. Aggarwal, Z. Kai, and W. Wei. 2018. NetWalk: A Flexible Deep Embedding Approach for Anomaly Detection in Dynamic Networks. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2672–2681. https://doi.org/10.1145/3219819.3220024

[56] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar. 2018. Efficient GAN-Based Anomaly Detection. In *The Workshop on International Conference on Learning Representations*. http://arxiv.org/pdf/1802.06222

[57] Y. L. Zhang, L. Li, J. Zhou, X. Li, and Z. H. Zhou. 2018. Anomaly Detection with Partially Observed Anomalies. In *WWW: International World Wide Web Conference*. 639–646. https://doi.org/10.1145/3184558.3186580

[58] Y. J. Zheng, X. H. Zhou, W. G. Sheng, Y. Xue, and S. Y. Chen. 2018. Generative Adversarial Network Based Telecom Fraud Detection at The Receiving Bank. *Neural Networks* 102 (2018), 78–86. https://doi.org/10.1016/j.neunet.2018.02.015

[59] C. Zhou and R. C. Paffenroth. 2017. Anomaly Detection with Robust Deep Autoencoders. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 665–674. https://doi.org/10.1145/3097983.3098052

[60] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh. 2019. AnomalyNet: An Anomaly Detection Network for Video Surveillance. *IEEE Transactions on Information Forensics and Security* 14, 10 (2019), 2537–2550. https://doi.org/10.1109/TIFS.2019.2900907

[61] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen. 2018. Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. In *International Conference on Learning Representations*. https://iclr.cc/Conferences/2018/Schedule?showEvent=12