Singapore Management University

Institutional Knowledge at Singapore Management University

Research Collection School Of Computing and Information Systems | School of Computing and Information Systems

9-2023

RoSAS: Deep semi-supervised anomaly detection with contamination-resilient continuous supervision

Hongzuo XU

Yijie WANG

Guansong PANG
Singapore Management University, gspang@smu.edu.sg

Songlei JIAN

Ning LIU

See next page for additional authors

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research

 Part of the Artificial Intelligence and Robotics Commons, and the Databases and Information Systems Commons

Citation

XU, Hongzuo; WANG, Yijie; PANG, Guansong; JIAN, Songlei; LIU, Ning; and WANG, Yongjun. RoSAS: Deep semi-supervised anomaly detection with contamination-resilient continuous supervision. (2023). *Information Processing and Management*. 60, (5), 1-17.
Available at: https://ink.library.smu.edu.sg/sis_research/8267

This Journal Article is brought to you for free and open access by the School of Computing and Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Computing and Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email cherylds@smu.edu.sg.

Author

Hongzuo XU, Yijie WANG, Guansong PANG, Songlei JIAN, Ning LIU, and Yongjun WANG

This journal article is available at Institutional Knowledge at Singapore Management University: https://ink.library.smu.edu.sg/sis_research/8267

RoSAS: Deep Semi-supervised Anomaly Detection with Contamination-resilient Continuous Supervision

Hongzuo Xu[a,b], Yijie Wang[a,b,*], Guansong Pang[c], Songlei Jian[b], Ning Liu[d], Yongjun Wang[b]

[a] *National Key Laboratory of Parallel and Distributed Computing*
[b] *College of Computer, National University of Defense Technology, Changsha Hunan 410073, China*
[c] *School of Computing and Information Systems, Singapore Management University, Singapore 178902, Singapore*
[d] *College of Information and Communications, National University of Defense Technology, Wuhan Hubei 430010, China*

arXiv:2307.13239v1 [cs.LG] 25 Jul 2023

Abstract

Semi-supervised anomaly detection methods leverage a few anomaly examples to yield drastically improved performance compared to unsupervised models. However, they still suffer from two limitations: 1) unlabeled anomalies (i.e., anomaly contamination) may mislead the learning process when all the unlabeled data are employed as inliers for model training; 2) only discrete supervision information (such as binary or ordinal data labels) is exploited, which leads to suboptimal learning of anomaly scores that essentially take on a continuous distribution. Therefore, this paper proposes a novel semi-supervised anomaly detection method, which devises *contamination-resilient continuous supervisory signals*. Specifically, we propose a mass interpolation method to diffuse the abnormality of labeled anomalies, thereby creating new data samples labeled with continuous abnormal degrees. Meanwhile, the contaminated area can be covered by new data samples generated via combinations of data with correct labels. A feature learning-based objective is added to serve as an optimization constraint to regularize the network and further enhance the robustness w.r.t. anomaly contamination. Extensive experiments on 11 real-world datasets show that our approach significantly outperforms state-of-the-art competitors by 20%-30% in AUC-PR and obtains more robust and superior performance in settings with different anomaly contamination levels and varying numbers of labeled anomalies. The source code is available at https://github.com/xuhongzuo/rosas/.

Keywords: Anomaly detection, Anomaly contamination, Continuous supervision, Semi-supervised learning, Deep learning

1. Introduction

Anomaly detection is to identify exceptional data objects that are deviated significantly from the majority of data, which has wide applications in many vital domains, e.g., network security, financial surveillance, risk management, and AI medical diagnostics (Pang et al., 2021b). Anomaly detection is often posited as an unsupervised problem due to the difficulty of accessing adequate labeled data (Han et al., 2022; Jiang et al., 2023). The past decade has witnessed a plethora of unsupervised anomaly detection methods that estimate/learn data normality via various data characteristics (e.g., proximity, probability, or clustering membership) or deep models (e.g., different kinds of Autoencoders or generative adversarial networks). However, these unsupervised methods often have many false alarms which can overwhelm human analysts, leading to the failure of investigating real threats. It is challenging, if not impossible, to accurately detect true anomalies of real interest without any prior information indicating what kind of data are anomalies.

In fact, in many real-world applications, there are often a few readily accessible anomaly examples. For example, some abnormal events such as credit card frauds or insiders' unauthorized access are reported (by users) or logged

*Corresponding author
Email addresses: xuhongzuo13@nudt.edu.cn (Hongzuo Xu), wangyijie@nudt.edu.cn (Yijie Wang), gspang@smu.edu.sg (Guansong Pang), jiansonglei@nudt.edu.cn (Songlei Jian), liuning17a@nudt.edu.cn (Ning Liu), wangyongjun@nudt.edu.cn (Yongjun Wang)

(in the system). Small genuine anomaly data can be directly retrieved from these records, without requiring extra annotations. This naturally inspires us to harness these true anomalies in combination with unlabeled data when training detection models. This learning paradigm falls into the category of semi-supervised learning (Chen et al., 2019; Kang et al., 2021; Van Engelen & Hoos, 2020; Yu et al., 2018) that permits using small labeled data as well as a large amount of unlabeled data. Recently, with the help of dozens of anomaly examples, semi-supervised methods have shown drastically improved detection performance compared to unsupervised methods that work on unlabeled data only (Ding et al., 2021, 2022; Jiang et al., 2023; Pang et al., 2018, 2019, 2023; Zhou et al., 2021, 2022).

By summarizing prior arts, this paper first proposes a general deep semi-supervised anomaly detection framework by introducing a two-stage network structure and a general learning objective. This framework presents a unifying view of this research line. More importantly, this framework reveals the following two key limitations of existing deep semi-supervised anomaly detection models that we aim to address in this study:

Robustness w.r.t. anomaly contamination. Many studies (Pang et al., 2018; Ruff et al., 2020; Wu et al., 2021; Zhou et al., 2021) assume all the unlabeled data as normal since anomalies are rare events. However, some anomalies are still hidden in the unlabeled set (i.e., *anomaly contamination*). This contamination might disturb anomaly detection models and blur the boundaries of normal patterns, leading to the potential overfitting problem. Some attempts (Pang et al., 2019, 2021a, 2023) have been made to address this problem by using a Gaussian prior when defining optimization targets or using concatenated data pairs as augmented training data.

Continuous supervision of anomaly score optimization. Anomaly detection models are typically required to output anomaly scores to indicate the degree of being abnormal for human investigation of the top-ranked anomalies. However, current models only use discrete supervision information, e.g., binary optimization targets (Pang et al., 2018, 2019, 2021a; Ruff et al., 2020; Wu et al., 2021; Zhou et al., 2021) or ordinal class labels (Pang et al., 2020, 2023), to optimize anomaly scores that essentially take on a continuous distribution. The lack of continuous supervision may result in suboptimal learning of anomaly scores. To the best of our knowledge, we are the first to raise this problem in anomaly detection.

To exemplify the issues described above, we use a toy dataset[1] in Figure 1. Figure 1 (a) visualizes the data with ground-truth annotations, in which the left panel uses the two most relevant dimensions as coordinate axes and the right panel is the T-SNE (Van der Maaten & Hinton, 2008) result. Most existing models use the contaminated discrete supervisory signals directly supplied by raw labels of the semi-supervised setting, as shown in Figure 1 (b). Data samples in this supervision are labeled by discrete values, and more importantly, this supervision is biased by unlabeled anomalies, i.e., anomaly contamination (e.g., two gray triangles highlighted in the blue rectangle). This supervision is not indicative enough to support the detection of the hard anomalies that are mixed up with inliers, or similar to the unlabeled anomalies. As shown in Figure 1 (d), five current state-of-the-art semi-supervised detectors suffer from these issues and fail to yield satisfactory detection results.

To fill these gaps, this paper further proposes a novel Robust deep Semi-supervised Anomaly Scoring method (termed RoSAS), in which the produced anomaly scores are optimized via *contamination-resilient continuous supervisory signals*. RoSAS follows our general network structure consisting of a feature representation module and an anomaly scoring module to directly yield anomaly scores, where the whole process is optimized in an end-to-end manner. Specifically, we first propose a mass interpolation method to diffuse the abnormality of labeled anomaly examples to the unlabeled area, which yields augmented data samples. As the interpolation process is measurable according to the diffusion intensity, these newly created data can be labeled with continuous values that faithfully indicate their abnormal degrees, thereby offering continuous supervision to straightly optimize the anomaly scoring mechanism. The located area of anomaly contamination can be covered by the new data generated by the interpolation of data combinations with correct labels. Even if the anomalies hidden in unlabeled data are used in interpolation, their negative effects can be diluted when they are grouped with genuine normal data or real anomalies in a mass. Consequently, new supervisory signals can better tolerate the anomaly contamination problem. Besides, our optimization process encourages the consistency between the anomaly score of each augmented sample and the score interpolation of their corresponding original data samples. This consistency learning can produce smoother anomaly scores to describe continuous abnormal degrees better. Additionally, we pose a feature learning-based objective that

[1]This toy dataset is generated via the `make_classification` function of the `Scikit-learn` library (Pedregosa et al., 2011). The dataset is described by ten features, including three informative features, five redundant features (i.e., random linear combinations of the informative features), and two noisy features. The anomaly class contains three clusters.

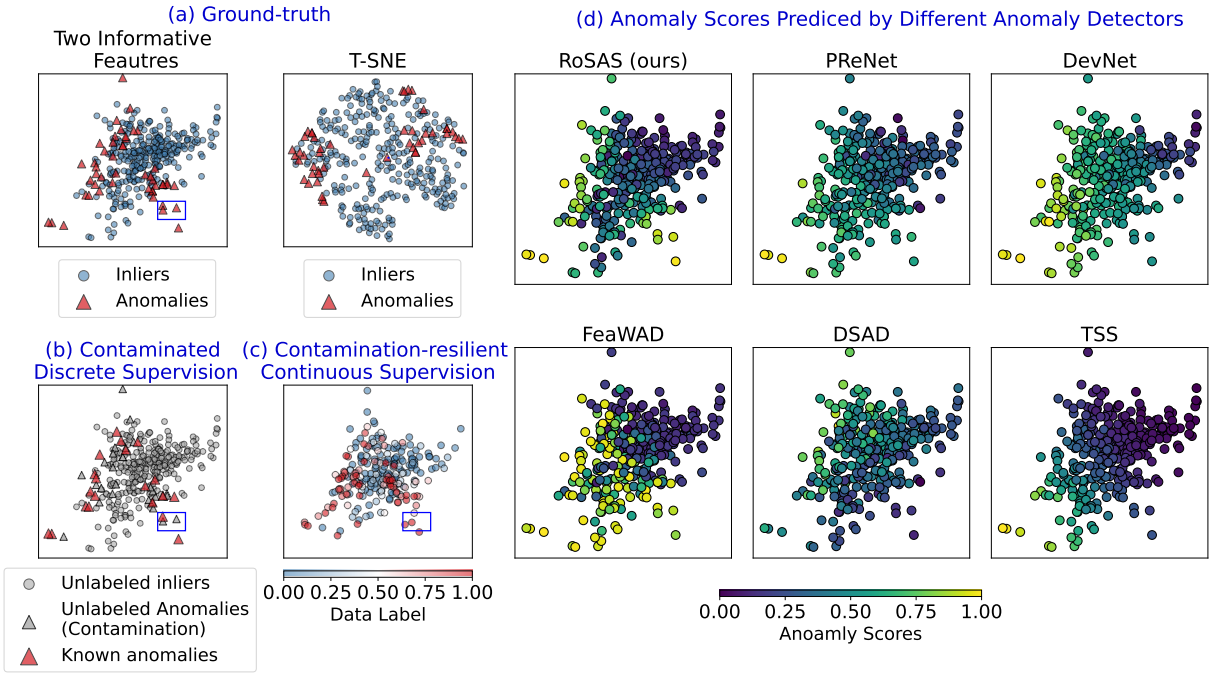


Figure 1: (**a**) Ground-truth labels of a toy case (the left panel uses two raw features that are informative to show the data distribution, and the right panel shows the 2-D data transformed by T-SNE). In the following sub-figures, we rely on the two raw informative features to visualize. (**b**) Raw supervision information (i.e., *contaminated discrete supervision*) directly offered by the semi-supervised setting. (**c**) *Contamination-resilient continuous supervision* generated by our model. (**d**) Anomaly scoring results of our method RoSAS vs. existing approaches including PReNet (Pang et al., 2023), DevNet (Pang et al., 2019, 2021a), FeaWAD (Zhou et al., 2021), DSAD (Ruff et al., 2020), and TSS (Zhang et al., 2017). The blue rectangle in (a)(b)(c) is used to highlight two real anomalies that are hidden in the unlabeled set (i.e., anomaly contamination). These two noisy points may mislead the learning model, but they are effectively covered in our supervision. The generated augmented samples in this area are labeled with higher values clearly indicating the anomalism of this field. Benefiting from the proposed contamination-resilient continuous supervision in (c), our method RoSAS produces more accurate anomaly scores than prior arts as shown in (d).

ensures effective isolation of labeled anomalies in the intermediate representation, which serves as an optimization constraint to further regularize the network and enhance the robustness w.r.t. anomaly contamination.

Figure 1 (c) illustrates the devised contamination-resilient continuous supervision, which is not only noise-tolerant but very faithful to the ground truth, demonstrating significantly higher supervision quality. Therefore, as depicted in Figure 1 (d), RoSAS produces more reliable anomaly scoring results than competing methods that rely on raw supervision information.

Our main contributions are summarized as follows.

- Motivated by the two limitations manifested by the general framework of this research line, we propose a novel semi-supervised anomaly detection method RoSAS, in which we devise a new kind of contamination-resilient continuous supervisory signals to optimize anomaly scores in an end-to-end manner.
- We propose a mass interpolation method in RoSAS to generate augmented data samples together with continuous values as data labels. In addition to offering continuous supervision, the created supervisory signals can tolerate anomaly contamination.
- We introduce consistency learning which encourages RoSAS to produce smoother anomaly scores, thus better describing abnormal degrees. We also set a feature learning-based objective to regularize RoSAS. The intermediate representation is constrained to further enhance its robustness w.r.t. anomaly contamination.

Extensive experiments show that: 1) RoSAS achieves significant AUC-PR and AUC-ROC improvement over state-of-the-art semi-supervised anomaly detection methods; 2) RoSAS obtains more robust and superior performance in

settings with different anomaly contamination levels and varying numbers of labeled anomalies. We also empirically show the advantage of the proposed contamination-resilient continuous supervisory signals over discretized, conventional ones and validate contributions of the consistency constraint in anomaly scoring and the regularizer based on feature learning.

2. Related Work

This section first reviews unsupervised anomaly detection and summaries semi-supervised models that exploit labeled anomaly examples.

2.1. Unsupervised Anomaly Detection

Traditional unsupervised anomaly detection identifies anomalies according to different data characteristics like proximity and probability (Bandaragoda et al., 2018; Li et al., 2020; Liu et al., 2008). The burgeoning of deep learning has fueled a plethora of deep anomaly detectors. In this research line, many studies (Ding et al., 2019; Gong et al., 2019; Lv et al., 2023; Xu et al., 2019; Zhang et al., 2019) train Autoencoders or generative adversarial networks to reconstruct/generate the original inputs. Self-supervised methods (Golan & El-Yaniv, 2018; Shenkar & Wolf, 2022; Xu et al., 2023b) define data-driven supervision and proxy tasks. These methods essentially learn intrinsic patterns of training data that are dominated by normal data, and loss values are directly used to estimate abnormal degrees during inference. In addition, some studies enhance traditional models by harnessing the strong representation capability of deep learning. Deep SVDD (Ruff et al., 2018) is based on support vector data description (Tax & Duin, 2004), and DIF (Xu et al., 2023a) enhances the isolation process of (Liu et al., 2008) by proposing deep representation ensemble. Basic insights in mainstream deep anomaly detectors can be also achieved via non-deep models. The literature (Xu et al., 2021b) uses tree models to realize the reconstruction pipeline. Although these unsupervised methods are intuitive and practical, without knowing real anomalies, they often lead to many false alarms which may overwhelm anomalies of real interest.

2.2. Semi-supervised Anomaly Detection

In contrast, relatively few studies consider semi-supervised anomaly detection utilizing limited anomaly examples. In this category, we also review the related literature that uses both labeled normal data and labeled anomalies since they can also work under this scenario by treating unlabeled data as normal.

The study (Zhang et al., 2018b) employs canonical clustering to divide labeled anomalies into k clusters and detect anomalies by a $(k+1)$-class classifier. Non-deep unsupervised anomaly detection methods can be also enhanced to leverage weak incomplete supervision. Barbariol & Susto (2022) extend ensemble-based isolation forest (Liu et al., 2008) by leveraging supervision information to filter ensemble members, which improves detection performance and simultaneously reduces computational costs.

This incomplete supervision can be also leveraged in deep models to learn a good representation. Some methods map input data to a representation space and explicitly impose specific criteria such as triplet loss (Pang et al., 2018) and anomaly-informed one-class loss (Ruff et al., 2020) upon the representation. They further employ distance-based anomaly scoring protocols upon this learned representation space. Besides, data representations can be also implicitly learned via Autoencoders or generative adversarial networks. Huang et al. (2020) propose a novel encoder-decoder-encoder structure. It modifies the reconstruction loss to force the network to reconstruct labeled anomalies to pre-defined noises. Bidirectional GAN is used in (Tian et al., 2022), in which labeled anomalies are used to learn a probability distribution, and the distribution can assign low-density values to labeled anomalies. These methods are indirectly optimized to yield abnormal degrees of data samples, and anomaly scores can be only obtained in an isolated manner.

Some advanced deep approaches are in an end-to-end fashion to directly optimize the produced anomaly scores. The pioneering work in this research line (Pang et al., 2019, 2021a) assumes anomaly scores of normal data follow a Gaussian distribution and yield the reference score. It further employs the z-score function to define the deviation loss to ensure anomaly scores of labeled anomalies significantly deviate from the reference. An Autoencoder is added to the above framework in (Zhou et al., 2021). In addition to a deviation loss imposed on the derived anomaly scores, the reconstruction error of labeled anomalies is optimized to be as larger as a pre-defined margin. By defining the

ordinal target of paired data samples, Pang et al. (2019) use mean absolute error to optimize anomaly scores. The cross-entropy loss is used in (Ding et al., 2022) to classify labeled anomalies, transferred pseudo anomalies, and latent residual anomalies from unlabeled data.

It is also noteworthy that, except for tabular data or images, related studies also consider this semi-supervised learning paradigm of anomaly detection in graph data (Ding et al., 2021; Dou et al., 2020; Zhou et al., 2022) and time series (Carmona et al., 2022; Huang et al., 2022).

3. A General Framework of Deep Semi-supervised Anomaly Detection

Problem Statement. We assume a few labeled anomaly examples $\mathcal{X}_A$ are accessible in addition to large-scale unlabeled training data $\mathcal{X}_U$, where $|\mathcal{X}_A| \ll |\mathcal{X}_U|$, i.e., the quantity of labeled anomalies is very small compared to the number of true anomalies and the whole dataset. Given the training data $\mathcal{X} = \mathcal{X}_U \cup \mathcal{X}_A$, an anomaly detection model is trained to assign higher scores to data samples with higher likelihoods to be anomalies.

3.1. General Framework

We below introduce a general framework of deep semi-supervised anomaly detection, and this framework can well cover representative existing models (Carmona et al., 2022; Pang et al., 2018, 2019, 2021a, 2023; Ruff et al., 2020; Wu et al., 2021; Zhou et al., 2021) and summarize their limitations.

We first define the network structure of the framework. Let $f : \mathcal{X} \mapsto \mathbb{R}$ represent the network that outputs anomaly scores given the input data $\mathcal{X}$. The whole procedure can be divided into a feature representation module $\phi : \mathcal{X} \mapsto \mathbb{R}^H$ and an anomaly scoring module $\psi : \mathbb{R}^H \mapsto \mathbb{R}$. Feature representation module ϕ aims to map $\mathcal{X}$ into a feature space with dimensionality H. Anomaly scoring module ψ outputs final anomaly scores based on the intermediate representation. Anomaly detection network f is denoted as:

$$f(\mathbf{x}) = \psi(\phi(\mathbf{x}; \Theta_\phi); \Theta_\psi), \tag{1}$$

where Θ_ϕ and Θ_ψ are network parameters in ϕ and ψ.

We then define a general learning objective. Under the semi-supervised setting, each data sample in the training set can be assigned a target. Let $\mathcal{D} = \{(\mathbf{x}, y) \in \mathcal{X} \times \mathcal{Y}\}$ with $\mathcal{Y} = \{y^+, y^-\}$ be a set of training samples, where y^+ denotes labeled anomalies and y^- denotes unlabeled data. Although most of $(\mathbf{x}, y^-)$ are genuine normal samples, there are still some unlabeled anomalies that are wrongly assigned y^-. Data augmentation techniques can be also used to obtain a new training set $\tilde{\mathcal{D}} = \{(\tilde{\mathbf{x}}, \tilde{y})\}$. A general objective function is defined as follows:

$$\begin{aligned} \min_{\{\Theta_\phi, \Theta_\psi\}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\Big[\mathcal{L}_D(\psi(\phi(\mathbf{x})), y)\Big] + \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\Big[\mathcal{L}'_D(\phi(\mathbf{x}), y)\Big] \\ + \mathbb{E}_{(\tilde{\mathbf{x}},\tilde{y})\sim\tilde{\mathcal{D}}}\Big[\mathcal{L}_{\tilde{D}}(\psi(\phi(\tilde{\mathbf{x}})), \tilde{y})\Big] + \mathbb{E}_{(\tilde{\mathbf{x}},\tilde{y})\sim\tilde{\mathcal{D}}}\Big[\mathcal{L}'_{\tilde{D}}(\phi(\tilde{\mathbf{x}}), \tilde{y})\Big]. \end{aligned} \tag{2}$$

The above equation can be interpreted as the optimization of the representation $\phi(\cdot)$ and/or the final anomaly scores $\psi(\phi(\cdot))$ by using supervision signals provided by original data $\mathcal{D}$ and/or augmented data $\tilde{\mathcal{D}}$.

3.2. Generalization of Current Studies

As for the network structure in Eqn (1), different network structures are used according to data types and data characteristics, e.g., multi-layer perceptron net is used for multi-dimensional tabular data (Pang et al., 2018, 2019, 2021a, 2023; Wu et al., 2021; Zhou et al., 2021), convolutional net is used for image data (Ruff et al., 2020), and temporal net is used for time series (Carmona et al., 2022).

The proposed objective function Eqn. (2) can well cover existing deep semi-supervised anomaly detectors by specifying each of its terms, as shown in Table 1. We below explain their instantiation method in detail.

- Deep SAD (Ruff et al., 2020) defines $\mathcal{L}'_D$. Upon the representation space, labeled anomalies are repulsed to be distant to a pre-defined center $\mathbf{c}$ as far as possible, and unlabeled data are expected to be included in a compact hypersphere with the minimum volume taking $\mathbf{c}$ as the center.

- FeaWAD (Zhou et al., 2021) first instantiates $\mathcal{L}_D$. It is optimized to enlarge anomaly scores of labeled anomalies to a pre-defined margin e and maps scores of unlabeled data to zero. $\mathcal{L}'_D$ is further instantiated by a reconstruction loss with the help of an Autoencoder structure.
- DevNet (Pang et al., 2019, 2021a) specifies $\mathcal{L}_D$. It proposes a z-score-based deviation function by assuming a pre-defined Gaussian prior of anomaly scores and sampling reference scores μ and standard deviation values σ from this distribution.
- PReNet (Pang et al., 2023) specifies $\mathcal{L}_{\tilde{D}}$ as Mean Absolute Error (MAE) between the scores of concatenated pairs (anomaly-unlabeled, anomaly-anomaly, and unlabeled-unlabeled) and pre-defined ordinal regression targets (e_1, e_2, and e_3).

REPEN (Pang et al., 2018) and NCAD (Carmona et al., 2022) also repulse labeled anomalies upon the representation space, as has been done in Deep SAD. PLSD (Wu et al., 2021) is similar to PReNet by replacing MAE loss with cross-entropy loss. Therefore, these methods are omitted in Table 1.

Table 1: Instantiation method and gaps of existing deep semi-supervised anomaly detection studies

Anomaly Detectors	$\mathcal{L}_D$ / $\mathcal{L}_{\tilde{D}}$: **anomaly score optimization**	$\mathcal{L}'_D$ / $\mathcal{L}'_{\tilde{D}}$: **representation optimization**	**Robustness**	**Continuous supervision**
Deep SAD (Ruff et al., 2020)	-	$\mathbb{1}_{y^+}\|\phi(\mathbf{x})-\mathbf{c}\|^{-1}+$ $\mathbb{1}_{y^-}\|\phi(\mathbf{x})-\mathbf{c}\|$	✗	✗
FeaWAD (Zhou et al., 2021)	$\mathbb{1}_{y^+}\max(0, e-\psi(\phi(x)))+$ $\mathbb{1}_{y^-}\|\psi(\phi(\mathbf{x}))\|$	$\mathbb{1}_{y^+}\max(0, e-\|\phi(\mathbf{x})-\mathbf{x}\|)+$ $\mathbb{1}_{y^-}\|\phi(\mathbf{x})-\mathbf{x}\|$	✗	✗
DevNet (Pang et al., 2019, 2021a)	$\mathbb{1}_{y^+}\max(0, e-\frac{\psi(\phi(\mathbf{x}))-\mu}{\sigma})+$ $\mathbb{1}_{y^-}\|\frac{\psi(\phi(\mathbf{x}))-\mu}{\sigma}\|$	-	✓	✗
PReNet (Pang et al., 2023)	$\mathbb{1}_{\langle y_i^+, y_j^+\rangle}\|\psi(\phi(\langle\mathbf{x}_i,\mathbf{x}_j\rangle))-e_1\|+$ $\mathbb{1}_{\langle y_i^+, y_j^-\rangle}\|\psi(\phi(\langle\mathbf{x}_i,\mathbf{x}_j\rangle))-e_2\|+$ $\mathbb{1}_{\langle y_i^-, y_j^-\rangle}\|\psi(\phi(\langle\mathbf{x}_i,\mathbf{x}_j\rangle))-e_3\|$	-	✓	✗
RoSAS (ours)	$\ell(\psi(\phi(\tilde{\mathbf{x}})), \tilde{y})+$ $\ell(\psi(\phi(\tilde{\mathbf{x}})), \sum_{i=1}^{k}\lambda_i\psi(\phi(\mathbf{x}_i)))$	$\max(0, e+d(\phi(\mathbf{x}^-), \phi(\mathbf{q})))-$ $d(\phi(\mathbf{x}^+), \phi(\mathbf{q}))$	✓	✓

3.3. Limitations of Current Studies

By looking into Table 1, we perceive two key gaps in these existing approaches, i.e., *robustness w.r.t. contamination* and *continuous supervision of optimization*.

3.3.1. Robustness w.r.t. contamination

Deep SAD and FeaWAD use unlabeled data as an opposite data class against labeled anomalies. They define a specific loss term (starting with $\mathbb{1}_{y^-}$) to *indistinguishably* map all of these unlabeled data to a unified target. This operation seems to be reasonable due to the unsupervised nature of anomaly detection (i.e., anomalies are rare data). To further enhance detection performance, we need to consider the negative effect brought by anomaly contamination in unlabeled data and improve the model robustness. DevNet (Pang et al., 2019, 2021a) assumes a Gaussian distribution prior of optimization targets. Due to the flexibility of Gaussian distribution, it can partially eliminate interference. PReNet (Pang et al., 2023) uses vector concatenation of data pairs to redefine three surrogate classes. This kind of data combination can resist small anomaly contamination since the interference of noisy samples can be mitigated when they are combined with genuine normal data or labeled anomalies.

3.3.2. Continuous supervision

Anomaly scores produced by anomaly detection models are expected to indicate abnormal degrees, and human investigators can examine the reported suspicious data in descending order of anomaly scores. Deep SAD, DevNet and FeaWAD utilize discrete binary supervision to respectively map labeled anomalies and unlabeled data to *two extremes* during training, but their models are required to output continuous anomaly scores during inference. Specifically, Deep SAD uses one fixed center $\mathbf{c}$ (unlabeled data are gathered at this center, and anomalies are repelled), FeaWAD directly maps anomaly scores to zero and a pre-defined margin e, and DevNet first calculates z-scores of anomaly

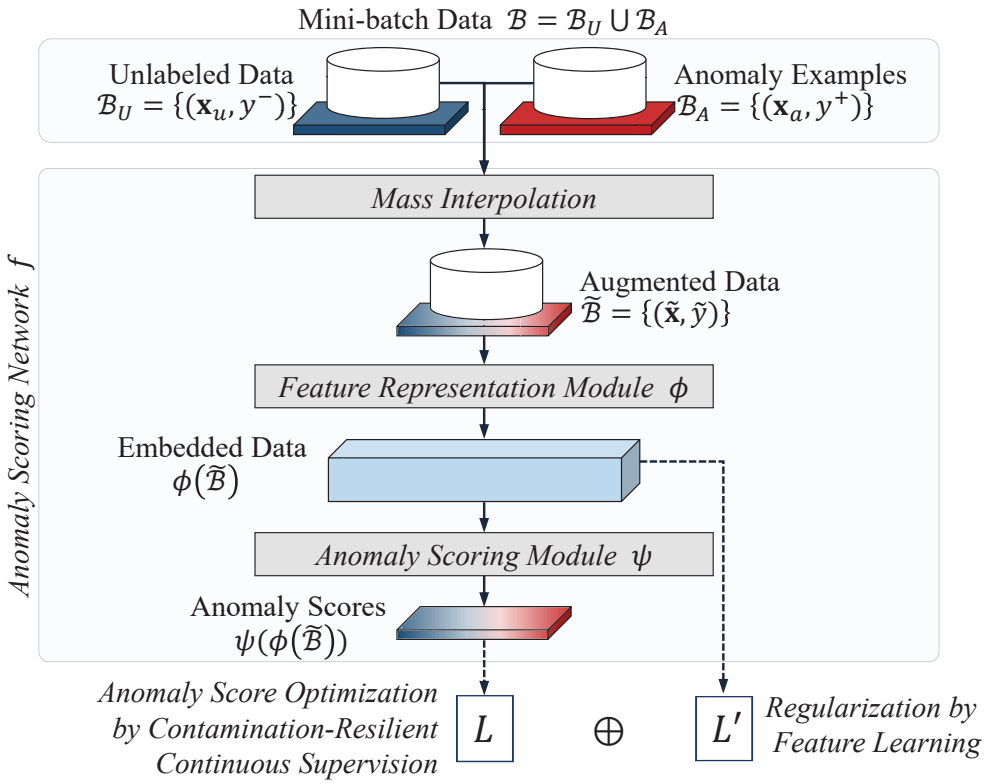


Figure 2: The overall procedure of RoSAS. Each training mini-batch $\mathcal{B}$ is composed of unlabeled data $\mathcal{B}_U$ and anomaly examples $\mathcal{B}_A$. The derived anomaly scores are end-to-end optimized by loss function L. L is defined based on contamination-resilient continuous supervision signals that are offered by augmented samples $\tilde{\mathcal{B}}$. A feature learning-based objective L' on the intermediate representation ϕ is added to further regularizes the network. L and L' are assembled via dynamic weight averaging $\oplus$.

scores and employs zero and a margin e as two extreme targets. Instead of using two extremes, PReNet employs three pre-defined ordinal targets, but this is also a kind of discrete supervision. Prior arts utilize the above discrete supervision information to optimize the continuously distributed anomaly scores. Due to the lack of continuous supervision, these models may fail to learn how to subtly describe abnormal degrees, resulting in suboptimal anomaly scoring mechanism.

4. The proposed RoSAS

This paper proposes a concrete deep semi-supervised anomaly detection method termed RoSAS. The overall procedure is shown in Figure 2. As described in Eqn. (1), RoSAS also follows the basic network structure f consisting of a feature representation module $\phi(\cdot|\Theta_\phi)$ and a scoring module $\psi(\cdot|\Theta_\psi)$. RoSAS is optimized by the loss function L and the regularizer L'.

The network architecture of ϕ and ψ is determined according to the input data types and/or data characteristics. In terms of the design of loss function L, the simplest way is to directly treat the whole unlabeled set as normal data, and discrete targets can be assigned to labeled anomaly examples and unlabeled data, as has been done in many prior studies (Pang et al., 2019; Zhou et al., 2021). However, these labels are inaccurate due to anomaly contamination and fail to sufficiently reflect anomaly scores that by definition take on a continuous distribution. It is very challenging to obtain reliable abnormal degrees of original training data since we do not exactly know whether one anomaly is more abnormal than another. Hence, we finally resort to synthesizing new data samples by diffusing the abnormality of these accessible labeled anomalies to the unlabeled area, and thus the abnormal degree is controllable. That is, the design of L is essentially to specify the term $\mathcal{L}_{\tilde{D}}$ in Eqn. (2). Specifically, based on the original training mini-batch

data $\mathcal{B}$, we propose a mass interpolation method to create a set of augmented data samples attached with continuous supervision targets $\tilde{\mathcal{B}} = \{(\tilde{\mathbf{x}}, \tilde{y})\}$. Note that this new supervision can also resist the anomaly contamination problem. The contaminated area can be covered by new data samples generated via combinations of data with correct labels. Anomaly contamination also becomes less harmful when these notorious unlabeled anomalies are combined with genuine normal samples or labeled anomalies during the interpolation. Consequently, RoSAS successfully devices contamination-resilient, continuous supervision of anomaly score optimization.

Further, motivated by the potential generalization and regularization effect of multi-task learning (Vandenhende et al., 2021), we define an additional objective L' upon the feature representation module ϕ to encourage significant deviations of labeled anomalies in the intermediate representation space. The network can be regularized by this new optimization constraint, which further improves the robustness to anomaly contamination.

Two loss terms L and L' are finally assembled via dynamic weight averaging $\oplus$ to avoid manually setting a fixed weight. Dynamic weight averaging $\oplus$ can balance the optimization pace at two loss terms.

We below present the design of the loss function L (Section 4.1), the regularization term L' (Section 4.2), and the dynamic averaging $\oplus$ (Section 4.3) in detail. We finally illustrate the procedure of RoSAS by giving its pseudo code (Section 4.4).

4.1. Anomaly Score Optimization by Contamination-resilient Continuous Supervision

RoSAS first produces new augmented data samples with controllable and reliable abnormal degrees via the mass interpolation method. Compared to directly using contaminated discrete targets, RoSAS can optimize anomaly scores as a regression problem with faithful continuous targets.

Specifically, based on the original mini-batch data $\mathcal{B} = \{(\mathbf{x}, y)\}$ with $y = 1$ for labeled anomalies and $y = -1$ for unlabeled data, RoSAS creates a novel mini-batch $\tilde{\mathcal{B}}$ of augmented data samples by the mass interpolation. These augmented data samples are synthesized as a weighted summation of k original data samples. Different weights of candidates $\{\lambda_1, \cdots, \lambda_k\}$ produce continuous targets in new supervision. $\tilde{\mathcal{B}}$ is defined as follows.

$$\tilde{\mathcal{B}} = \left\{(\tilde{\mathbf{x}}, \tilde{y}) | \tilde{\mathbf{x}} = \textstyle\sum_{i=1}^{k} \lambda_i \mathbf{x}_i, \tilde{y} = \sum_{i=1}^{k} \lambda_i y_i\right\}, \tag{3}$$

where $\sum_{i=1}^{k} \lambda_i = 1$, $\{(\mathbf{x}_i, y_i)\}_{i=1}^{k} \subset \mathcal{B}$, and λ_i is sampled from a continuous distribution.

As for the distribution of λ, inspired by (Zhang et al., 2018a), RoSAS uses Beta distribution, i.e., $\lambda \sim Beta(\alpha, \alpha)$. It is because adjusting distribution parameter α can produce different types of weights, e.g., a uniform distribution when $\alpha = 1$ or an approximate truncated normal distribution when α is a larger value. If $\alpha > 1$, interpolation weights will centralize around 0.5, and some noisy labels might be produced (e.g., mixing two anomalies may yield a new sample in the normal manifold, but an anomalous label is given to this new sample). This is also known as "manifold intrusion" (Guo et al., 2019). To tackle this problem, we use $\alpha = 0.5$ by default. In doing so, interpolation weights are more likely to be slightly larger/smaller than 0/1, which makes the interpolation located in the local regions of the original samples. Thus, these possible noisy labels can be reduced or eliminated.

The loss function L measures the empirical risks of derived anomaly scores of augmented samples compared to the continuous targets. Additionally, we add a consistency term to measure the difference between each augmented sample's anomaly score and the weighted summation of their original data instances' anomaly scores using the same interpolation weights. This consistency learning is to encourage the network to produce smoother anomaly scores, thus better describing abnormal degrees. Therefore, L is finally defined as:

$$L = \mathbb{E}_{(\tilde{\mathbf{x}}, \tilde{y}) \sim \tilde{\mathcal{B}}}\left[\ell\Big(\psi(\phi(\tilde{\mathbf{x}})), \tilde{y}\Big) + \ell\Big(\psi(\phi(\tilde{\mathbf{x}})), \sum_{i=1}^{k} \lambda_i \psi(\phi(\mathbf{x}_i))\Big)\right], \tag{4}$$

where $\{\mathbf{x}_i\}_{i=1}^{k}$ is the original data samples when creating $\tilde{\mathbf{x}}$ as defined in Eqn. (3), and $\ell(\cdot, \cdot)$ is a base regression loss.

The loss function L not only fulfills continuous optimization but can tolerate anomaly contamination. The unlabeled set is still dominated by genuine normal data because of the rarity of anomalies. The contaminated area can be calibrated via new data samples that are augmented from a group of data with correct labels. Even if these noisy unlabeled anomalies are sampled in Eqn. (3), they are likely to be combined with labeled anomalies or real normal data. That is, the generation process of augmented data samples also dilutes the anomaly contamination in a simple yet effective manner. Therefore, RoSAS is more robust w.r.t. anomaly contamination.

4.2. Regularization by Feature Learning

The feature representation module $\phi : \mathcal{X} \mapsto \mathbb{R}^H$ maps input data into a new feature space. We further define a new loss term L' upon this intermediate representation space, which serves as a new optimization constraint to regularize the network and further enhance the robustness.

To fully leverage these labeled anomalies, L' is designed to learn a feature representation that can effectively repulse these labeled anomaly examples from unlabeled data (the majority of unlabeled data is normal). Let $\mathbf{q}$ be an anchor data object, and we utilize the difference between the deviation of unlabeled-anchor and anomaly-anchor pairs to measure the separability of labeled anomalies, which is defined as follows.

$$L' = \mathbb{E}_{\substack{(\mathbf{x}^+, y^+)\sim\mathcal{B}_A \\ (\mathbf{x}^-, y^-)\sim\mathcal{B}_U}}\Big[\max\Big(d(\phi(\mathbf{x}^-), \phi(\mathbf{q})) - d(\phi(\mathbf{x}^+), \phi(\mathbf{q})) + e, 0\Big)\Big], \tag{5}$$

where $d(\cdot, \phi(\mathbf{q}))$ indicates the deviation given the anchor data, and e is a margin. Different distance functions or similarity measures can be used. Considering the simplicity, we employ Euclidean distance here. $\mathcal{B}_U$ and $\mathcal{B}_A$ are unlabeled data and labeled anomalies in mini-batch $\mathcal{B}$. In practical implementation, a mini-batch of anchor data is sampled from the unlabeled set along with mini-batch $\mathcal{B}$, i.e., $\mathbf{q} \in \mathcal{B}_q, \mathcal{B}_q \subset \mathcal{X}_u$. Anchor data can also be determined as representative normal prototypes if labeled normal data are available.

It is noteworthy that L' uses a relative and soft manner to judge whether these labeled anomalies are effectively separated by introducing a reference divergence degree $d(\phi(\mathbf{x}^-), \phi(\mathbf{q}))$ between unlabeled data and anchor data. It avoids blindly enlarging $d(\phi(\mathbf{x}^+), \phi(\mathbf{q}))$, i.e., the anomalies that have been successfully deviated are no longer required to be optimized; thus, the optimizer can focus on true errors. On the other hand, even if unlabeled anomalies are wrongly identified as anchor data $\mathbf{q}$ or $\mathbf{x}^-$ in Eqn. 5, this function can still work to isolate labeled anomalies.

4.3. Dynamic Averaging

Instead of setting a fixed weight, the loss term L and the regularizer L' are assembled via dynamic weight averaging (Liu et al., 2019), i.e.,

$$wL + (1 - w)L', \tag{6}$$

where w is defined according to the optimization pace (loss descending rate) of L and L'. w is defined as follows.

$$w = \frac{\exp(L/T\bar{L})}{\exp(L/T\bar{L}) + \exp(L'/T\bar{L}')}, \tag{7}$$

where $\bar{L}$ and $\bar{L}'$ are the average loss of the last training epoch, and T is the temperature as used in the softmax function.

4.4. Algorithm of RoSAS

Algorithm 1 presents the training procedure of RoSAS. Step 1 initializes the loss terms for the subsequent dynamic weight averaging. For each training batch, a mini-batch of known anomalies $\mathcal{B}_a$ of size b is sampled from $\mathcal{X}_A$, $2b$ data objects are sampled from $\mathcal{X}_U$ as mini-batch $\mathcal{B}_u$ and anchor $\mathcal{B}_q$ in Steps 4-5. Step 6 creates a mini-batch of augmented data. The scoring loss and the regularization term are computed in Steps 7-8. Dynamic weights are adjusted in Step 9. Step 10 performs back propagation to optimize the network parameters w.r.t. the loss $wL + (1 - w)L'$. Step 12 updates the average losses.

The computation of loss term L and L' has an overall time complexity of $O(n\_epoch * n\_batch * b * H)$, where H is the representation dimension. The time complexity of RoSAS also depends on the network structure. We take a multi-layer perceptron network with u-hidden layer as an example, the feed-forward propagation incurs $O(n\_epoch * n\_batch * b * (Dh_1 + h_1h_2 + \cdots + h_u * 1))$, where h_i is the number of hidden units in the i-th hidden layer.

5. Experiments

In this section, we first describe experimental setup (Section 5.1) and conduct experiments to answer the following questions:

- **Effectiveness**: Is RoSAS more effective than state-of-the-art anomaly detectors on real-world datasets? Can RoSAS handle different types of anomalies? (Section 5.2)

Algorithm 1 Training of RoSAS

Input: Labeled anomaly examples - $\mathcal{X}_A$, unlabeled data - $\mathcal{X}_U$
Output: Anomaly Scoring network - $\psi(\phi(\cdot;\Theta_\phi);\Theta_\psi)$

```
Initialize L̄ ← 1, L̄′ ← 1
for t = 1 to n_epoch do
    for j = 1 to n_batch do
        B_A ← randomly sample b data objects from X_A
        B_U, B_q ← randomly sample 2b data objects from X_U
        B̃ ← create augmented mini-batch by Eqn. (3)
        Compute loss L by Eqn. (4)
        Compute regularizer L′ by Eqn. (5)
        Compute weights w by Eqn. (7)
        Optimize parameters {Θ_φ, Θ_ψ} w.r.t. wL + (1 − w)L′
    end for
    L̄, L̄′ ← average loss over the current epoch
end for
return ψ(φ(·))
```

- **Robustness**: How does the robustness of RoSAS and its competitors when the unlabeled set is contaminated by different levels of anomalies? (Section 5.3)
- **Data Efficacy**: Can RoSAS fully leverage different numbers of labeled anomalies? (Section 5.4)
- **Scalability Test**: How does the time efficiency of RoSAS compared to its competitors? (Section 5.5)
- **Ablation Study**: Do key designs contribute to better anomaly detection performance? (Section 5.6)
- **Sensitivity**: How do the hyper-parameters influence the detection performance of RoSAS? (Section 5.7)

5.1. Experimental Setup

5.1.1. Datasets

Eleven publicly available real-world datasets are used[2]. The dataset information is reported in Table 2, including abbreviation (Abbr.), domain/task, data dimensionality (D), the number of data samples (N), and the anomaly ratio (δ). The first eight datasets are with real anomalies, which cover three important real-world applications of anomaly detection in cybersecurity, medicine, and finance. The last three datasets are from ODDS, a popular repository of anomaly detection datasets, and they contain semantic anomalies. All of these datasets are broadly used as benchmarks in many anomaly detection studies, e.g., (Bandaragoda et al., 2018; Pang et al., 2019; Xu et al., 2023a). We scale each feature to $[0, 1]$ via min-max normalization. All the datasets are separated by a random 60:20:20 train-valid-test split while maintaining the original anomaly proportion.

5.1.2. Competitors

We employ ten anomaly detection models from three categories as competing methods of RoSAS:

- *Semi-supervised Anomaly Detector*: Five deep semi-supervised anomaly detection methods including PReNet (Pang et al., 2023), FeaWAD (Zhou et al., 2021), DevNet (Pang et al., 2019, 2021a), Deep SAD (DSAD for short) (Ruff et al., 2020), and BiGAN (Tian et al., 2022) are used. TiWS-iForest (WSIF for short) (Barbariol & Susto, 2022) is an enhanced version of (Liu et al., 2008), which leverages weak supervision to improve detection performance. These competitors fall into different categories of existing techniques, representing the state-of-the-art performance of this semi-supervised setting.

[2]These datasets are available at https://github.com/GuansongPang/ADRepository-Anomaly-detection-datasets, https://www.unb.ca/cic/datasets/, and http://odds.cs.stonybrook.edu/

Table 2: Dataset information. Abbr. is the dataset abbreviation used in the following experiments. D and N denote data dimensionality and data size per dataset, respectively. δ indicates the anomaly ratio.

Data	**Abbr.**	**Domain/Task**	D	N	δ
CIC-DoHBrW2020	DoH	Intrusion Detection	29	1,167,136	21.4%
CIC-IDS2017 WebAttack	WebAttack	Intrusion Detection	78	700,284	0.3%
CIC-IDS2017 PortScan	PortScan	Intrusion Detection	78	816,385	19.5%
UNSW-NB15 Exploit	Exploit	Intrusion Detection	196	96,000	3.1%
UNSW-NB15 Backdoor	Backdoor	Intrusion Detection	196	95,329	2.4%
Thyroid disease	Thyroid	Disease Diagnosis	21	7,200	7.4%
KDD Cup 2014 Donors	Donars	Funding Prediction	10	619,326	5.9%
Credit card fraud detection	Fraud	Fraud Detection	29	284,807	0.2%
Covertype	Cover	Ecosystem	10	286,048	1.0%
Letter recognition	Letter	Recognition	32	1,600	6.3%
Pen-based recognition	Pendigits	Recognition	16	6,870	2.3%

- *PU learning-based Method*: Learning from positive and unlabeled data (PU learning) is also a related field if we treat anomalies as positive data. We choose a representative PU learning-based anomaly detector (Zhang et al., 2017) as our competitor, which combines the two-stage strategy and the cost-sensitive strategy (TSS for short).

- *Unsupervised Anomaly Detector*: DIF (Xu et al., 2023a), IF (Liu et al., 2008), and COP (Li et al., 2020) are employed. DIF is an isolation-based method that is empowered by deep representation ensemble. IF is a popular anomaly detection algorithm that is broadly used in many industrial applications, and COP is the latest probability-based approach. Note that they are only used as baselines to examine whether our method and other semi-supervised approaches obtain significantly improved performance.

5.1.3. Parameter Settings and Implementations

In RoSAS, the learning rate is set as 0.005, intermediate representation dimension H is 128. As for the parameters in the loss function, we use $k = 2$, $\alpha = 0.5$, and $e = 1$. Smooth-ℓ_1 loss function is adopted as the base regression loss in L. The batch size b is 32. RoSAS uses the Adam optimizer with an ℓ_2-norm weight decay regularizer. The temperature T in dynamic weight averaging is 2. RoSAS uses a multi-layer perceptron network structure since the used experimental datasets are multi-dimensional data. The representation module and the scoring module both adopt a one-hidden-layer structure. The number of hidden units in the representation network is set as $h_1 = D + \lfloor \frac{1}{2}(H - D) \rfloor$, and the scoring network uses $h_2 = \lfloor \frac{1}{2}H \rfloor$. We use LeakyReLU activation in the hidden layers and the tanh function to normalize final anomaly scores.

All the detectors are implemented in Python. The implementations of PReNet, DevNet, FeaWAD, DSAD, WSIF, and BiGAN are released by their original authors. The source code of TSS is publicly available. RoSAS, DSAD, and BiGAN employ the PyTorch framework, and PReNet, DevNet, and FeaWAD are based on Keras. We use implementations of COP and IF the `pyod` (Zhao et al., 2019) package.

5.1.4. Performance Evaluation Metrics and Computing Infrastructure

Following the popular experiment protocol of anomaly detection studies (Pang et al., 2019, 2023; Ruff et al., 2020; Xu et al., 2021a, 2023a), two performance evaluation metrics, i.e., the Area under the Precision-Recall Curve (AUC-PR) and the Area under the Receiver Operating Characteristic Curve (AUC-ROC), are used. ROC curve indicates true positives against false positives, while points in PR curve are pairs of precision value and recall value of the anomaly class given different thresholds. These two metrics range from 0 to 1. Higher values indicate better performance. AUC-PR is more practical in real-world applications because it directly relates to benefits and costs of detection results, and achieving high AUC-PR is more challenging. Therefore, we take AUC-PR as the main detection performance metric in the following experiments. We report the average AUC-PR and AUC-ROC scores on each dataset over ten independent runs. Additionally, we employ the paired *Wilcoxon* signed-rank test to determine if the AUC-ROC/AUC-PR of RoSAS and each of its contenders are significantly different. It can examine the statistical significance of the improvement of RoSAS against existing state-of-the-art performance.

Table 3: AUC-PR and AUC-ROC performance (± standard deviation) of RoSAS and its competing methods. The best performer is boldfaced.

	Data	RoSAS	PReNet	DevNet	FeaWAD	DSAD	TSS	WSIF	BiGAN	DIF	IF	COP
AUC-PR	DoH	**0.893**$_{\pm 0.005}$	0.712$_{\pm 0.021}$	0.628$_{\pm 0.005}$	0.741$_{\pm 0.068}$	0.842$_{\pm 0.017}$	0.610$_{\pm 0.002}$	0.546$_{\pm 0.024}$	0.561$_{\pm 0.055}$	0.396	0.427	0.340
	WebAttack	**0.781**$_{\pm 0.051}$	0.223$_{\pm 0.055}$	0.220$_{\pm 0.078}$	0.320$_{\pm 0.081}$	0.458$_{\pm 0.116}$	0.136$_{\pm 0.005}$	0.026$_{\pm 0.011}$	0.050$_{\pm 0.034}$	0.003	0.004	0.004
	PortScan	**0.999**$_{\pm 0.000}$	0.983$_{\pm 0.005}$	0.973$_{\pm 0.030}$	0.995$_{\pm 0.005}$	0.998$_{\pm 0.001}$	0.990$_{\pm 0.001}$	0.585$_{\pm 0.131}$	0.605$_{\pm 0.172}$	0.181	0.180	0.135
	Exploit	**0.740**$_{\pm 0.026}$	0.560$_{\pm 0.056}$	0.450$_{\pm 0.084}$	0.510$_{\pm 0.075}$	0.623$_{\pm 0.037}$	0.523$_{\pm 0.076}$	0.199$_{\pm 0.053}$	0.272$_{\pm 0.081}$	0.255	0.060	0.083
	Backdoor	0.877$_{\pm 0.021}$	0.882$_{\pm 0.005}$	**0.884**$_{\pm 0.015}$	0.793$_{\pm 0.095}$	0.666$_{\pm 0.036}$	0.860$_{\pm 0.018}$	0.437$_{\pm 0.143}$	0.405$_{\pm 0.152}$	0.394	0.052	0.069
	Thyroid	**0.839**$_{\pm 0.046}$	0.436$_{\pm 0.024}$	0.252$_{\pm 0.034}$	0.322$_{\pm 0.038}$	0.304$_{\pm 0.099}$	0.197$_{\pm 0.010}$	0.537$_{\pm 0.080}$	0.083$_{\pm 0.006}$	0.074	0.131	0.134
	Donars	**1.000**$_{\pm 0.000}$	0.973$_{\pm 0.015}$	0.999$_{\pm 0.003}$	0.998$_{\pm 0.005}$	0.999$_{\pm 0.001}$	0.982$_{\pm 0.007}$	0.780$_{\pm 0.093}$	0.873$_{\pm 0.075}$	0.115	0.238	0.242
	Fraud	**0.831**$_{\pm 0.003}$	0.803$_{\pm 0.008}$	0.808$_{\pm 0.004}$	0.596$_{\pm 0.261}$	0.438$_{\pm 0.109}$	0.800$_{\pm 0.005}$	0.523$_{\pm 0.073}$	0.664$_{\pm 0.099}$	0.335	0.328	0.270
	Cover	**0.983**$_{\pm 0.003}$	0.957$_{\pm 0.008}$	0.907$_{\pm 0.055}$	0.939$_{\pm 0.031}$	0.922$_{\pm 0.026}$	0.916$_{\pm 0.003}$	0.673$_{\pm 0.154}$	0.603$_{\pm 0.132}$	0.191	0.063	0.061
	Letter	**0.501**$_{\pm 0.026}$	0.332$_{\pm 0.052}$	0.153$_{\pm 0.077}$	0.270$_{\pm 0.047}$	0.069$_{\pm 0.027}$	0.163$_{\pm 0.032}$	0.140$_{\pm 0.043}$	0.067$_{\pm 0.005}$	0.099	0.084	0.061
	Pendigits	**0.995**$_{\pm 0.013}$	0.906$_{\pm 0.073}$	0.884$_{\pm 0.053}$	0.907$_{\pm 0.064}$	0.983$_{\pm 0.023}$	0.916$_{\pm 0.012}$	0.583$_{\pm 0.117}$	0.161$_{\pm 0.199}$	0.302	0.366	0.239
	Average	**0.858**$_{\pm 0.018}$	0.706$_{\pm 0.029}$	0.651$_{\pm 0.040}$	0.672$_{\pm 0.070}$	0.664$_{\pm 0.045}$	0.645$_{\pm 0.016}$	0.457$_{\pm 0.084}$	0.395$_{\pm 0.092}$	0.213	0.176	0.149
	p-value	-	0.002	0.003	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
AUC-ROC	DoH	**0.955**$_{\pm 0.001}$	0.885$_{\pm 0.004}$	0.884$_{\pm 0.008}$	0.905$_{\pm 0.017}$	0.932$_{\pm 0.007}$	0.871$_{\pm 0.001}$	0.753$_{\pm 0.030}$	0.825$_{\pm 0.056}$	0.693	0.674	0.676
	WebAttack	0.988$_{\pm 0.006}$	0.952$_{\pm 0.004}$	0.934$_{\pm 0.019}$	0.976$_{\pm 0.012}$	**0.989**$_{\pm 0.002}$	0.917$_{\pm 0.001}$	0.830$_{\pm 0.083}$	0.896$_{\pm 0.012}$	0.530	0.607	0.630
	PortScan	**0.999**$_{\pm 0.000}$	0.997$_{\pm 0.001}$	0.994$_{\pm 0.006}$	0.999$_{\pm 0.001}$	0.999$_{\pm 0.000}$	0.997$_{\pm 0.000}$	0.886$_{\pm 0.079}$	0.910$_{\pm 0.090}$	0.518	0.506	0.304
	Exploit	**0.962**$_{\pm 0.010}$	0.951$_{\pm 0.006}$	0.913$_{\pm 0.008}$	0.949$_{\pm 0.018}$	0.956$_{\pm 0.014}$	0.936$_{\pm 0.008}$	0.803$_{\pm 0.053}$	0.862$_{\pm 0.021}$	0.864	0.743	0.771
	Backdoor	**0.985**$_{\pm 0.004}$	0.952$_{\pm 0.002}$	0.971$_{\pm 0.006}$	0.978$_{\pm 0.004}$	0.965$_{\pm 0.009}$	0.970$_{\pm 0.005}$	0.816$_{\pm 0.128}$	0.906$_{\pm 0.018}$	0.915	0.753	0.791
	Thyroid	**0.989**$_{\pm 0.003}$	0.809$_{\pm 0.010}$	0.728$_{\pm 0.022}$	0.747$_{\pm 0.011}$	0.729$_{\pm 0.048}$	0.716$_{\pm 0.018}$	0.904$_{\pm 0.060}$	0.548$_{\pm 0.020}$	0.497	0.646	0.703
	Donars	**1.000**$_{\pm 0.000}$	0.999$_{\pm 0.000}$	1.000$_{\pm 0.000}$	1.000$_{\pm 0.000}$	1.000$_{\pm 0.000}$	0.999$_{\pm 0.000}$	0.986$_{\pm 0.009}$	0.984$_{\pm 0.014}$	0.780	0.891	0.846
	Fraud	0.977$_{\pm 0.003}$	0.967$_{\pm 0.008}$	0.971$_{\pm 0.002}$	0.973$_{\pm 0.006}$	0.952$_{\pm 0.015}$	**0.979**$_{\pm 0.001}$	0.966$_{\pm 0.004}$	0.933$_{\pm 0.012}$	0.960	0.962	0.966
	Cover	**1.000**$_{\pm 0.000}$	**1.000**$_{\pm 0.000}$	0.999$_{\pm 0.001}$	0.999$_{\pm 0.001}$	0.998$_{\pm 0.002}$	0.999$_{\pm 0.000}$	0.992$_{\pm 0.005}$	0.975$_{\pm 0.022}$	0.963	0.887	0.867
	Letter	**0.873**$_{\pm 0.031}$	0.822$_{\pm 0.033}$	0.663$_{\pm 0.106}$	0.770$_{\pm 0.057}$	0.489$_{\pm 0.078}$	0.707$_{\pm 0.025}$	0.691$_{\pm 0.033}$	0.546$_{\pm 0.024}$	0.650	0.608	0.528
	Pendigits	**1.000**$_{\pm 0.000}$	0.999$_{\pm 0.001}$	0.996$_{\pm 0.001}$	0.999$_{\pm 0.001}$	**1.000**$_{\pm 0.000}$	0.997$_{\pm 0.001}$	0.977$_{\pm 0.011}$	0.705$_{\pm 0.237}$	0.951	0.955	0.905
	Average	**0.975**$_{\pm 0.005}$	0.939$_{\pm 0.006}$	0.914$_{\pm 0.016}$	0.936$_{\pm 0.012}$	0.910$_{\pm 0.016}$	0.917$_{\pm 0.005}$	0.873$_{\pm 0.045}$	0.826$_{\pm 0.048}$	0.756	0.748	0.726
	p-value	-	0.005	0.005	0.008	0.017	0.007	0.001	0.001	0.001	0.001	0.001

All the experiments are executed at a workstation with Intel Xeon Silver 4210R CPU, a single NVIDIA TITAN RTX GPU, and 64 GB RAM.

5.2. *Effectiveness*

5.2.1. *Anomaly Detection Performance on Real-world Datasets*

Following (Pang et al., 2019, 2023; Wu et al., 2021; Zhou et al., 2021), we randomly select 30 true anomalies from the training data per dataset as anomaly examples and the remaining training data as the unlabeled set. RoSAS and its five contenders are trained on training sets and used to measure abnormal degrees of data samples in testing sets. Labels of testing sets are strictly unknown to anomaly detectors and are only employed in the evaluation phase. As has been done in (Pang et al., 2019, 2023; Zhou et al., 2021), we also execute controlled experiments w.r.t. anomaly contamination rate. Each dataset is pre-processed by removing/injecting anomalies such that anomalies account for 2% of the unlabeled set. Specifically, the injected anomaly examples are obtained by replacing the values of 5% random features of a randomly selected real anomaly with the corresponding feature values of another real anomaly. This presents a simple and effective way to guarantee the presence of diverse and genuine (or weakly augmented) anomalies in the unlabeled data. This pre-processing step can cancel out the influence of different contamination ratios such that the performance of these anomaly detectors is comparable across datasets from various domains. Please note that we also examine the performance w.r.t. a wide range of contamination ratios in the following experiment.

Table 3 shows the AUC-PR and the AUC-ROC performance of RoSAS and its competing methods. RoSAS achieves the best AUC-PR or AUC-ROC performance on all the datasets. According to the p-values in the *Wilcoxon* signed-rank test, RoSAS significantly outperforms its ten competitors w.r.t. both AUC-PR and AUC-ROC at the 98% confidence level. Averagely, RoSAS obtains a substantial performance leap (approximate 20%-30% AUC-PR improvement) over exiting state-of-the-art competing methods PReNet, DevNet, FeaWAD, DSAD, and TSS. WSIF is a non-deep method, which is inferior to these deep state-of-the-art semi-supervised methods on complicated real-world datasets. BiGAN is originally designed for images, and its performance on tabular data might be downgraded. Benefiting from a few labeled anomalies, the average AUC-ROC performance of many semi-supervised methods exceeds 0.9, and RoSAS still gains 4%-7% improvement over state-of-the-art competitors. The performance of unsupervised anomaly detectors DIF, IF, and COP is distinctly inferior to all the semi-supervised approaches, which validates the importance of fully exploiting these readily accessible anomaly examples in real-world applications.

RoSAS achieves substantially superior detection performance with the help of the proposed contamination-resilient continuous supervision and feature learning-based regularization. The robustness of RoSAS is enhanced to better exploit the contaminated unlabeled set. RoSAS is also with direct fine-grained guidance to optimize anomaly scores more accurately. Therefore, RoSAS can better leverage dozens of anomaly examples and large-scale unlabeled data, resulting in effective semi-supervised anomaly detection. Note that PReNet obtains relatively better performance because it can resist small anomaly contamination thanks to its data combination operation. Other competitors do not consider the interference from noisy hidden anomalies and treat the whole unlabeled set as normal data. Also, all of these competing methods are only optimized by discrete supervision information that fails to indicate continuous abnormal degrees, resulting in suboptimal learning of anomaly scores.

5.2.2. Capability of Handling Different Types of Anomalies

We further investigate whether RoSAS can identify different types of anomalies. Anomalies can be classified into *clustered anomalies* and *scattered anomalies* according to the intra-class proximity (Xu et al., 2019; Zhou et al., 2022). Clustered anomalies (e.g., diseases and fraudulent activities) share similar behaviors, while scattered anomalies (e.g., exceptions in industrial systems) randomly appear out of the inlier distribution and have weak or even no connections with other individual samples. Besides, in the semi-supervised setting, there might be some *novel anomalies* that are different from labeled anomalies that appear during training. Novel anomalies are critical in real-world applications; for example, some advanced new attacks may pose severe threats to network security, but they are very different from those known intrusions. Due to the difficulty of knowing specific anomaly types in real-world datasets, we create three synthetic cases to validate the capability of handling these anomaly types. Training and testing data distributions of these three cases are demonstrated in Figure 3. Case 1 and Case 2 respectively contain clustered anomalies and scattered anomalies, and there is a cluster of novel anomalies in the testing set of Case 3.

Figure 3 further illustrates the detection results of RoSAS. By setting the threshold according to the size of true anomalies, we report both predicted anomaly scores and corresponding binary labels. We respectively analyze the detection results of the three cases below.

Case 1. In terms of clustered anomalies in Case 1, although the abnormal region is contaminated by unlabeled anomalies that are used as normal data, this region can be covered by new data samples that are augmented by our mass interpolation method. Labeled anomalies are over-sampled during training, and the unlabeled anomalies are still rare compared to genuine normal data because anomalies themselves are rare. The contamination can be corrected by new data generated via the interpolation of data combinations with correct labels, and thus RoSAS can effectively identify these clustered anomalies during inference.

Case 2. As for scattered anomalies in Case 2, unlabeled anomalies may not largely influence the training process. However, one key issue of this case is the "manifold intrusion" problem. For instance, the interpolation into labeled anomalies may create augmented data samples in the normal distribution, but they are labeled by high abnormal degrees. To alleviate this problem, RoSAS uses Beta distribution with $\alpha = 0.5$ in the mass interpolation process, thereby making most interpolation located in the local regions of original samples. This may still raise an inevitable limitation. Namely, RoSAS gives slightly higher anomaly scores to some margin points of the normal manifold, and there are two false positives as shown in the binary prediction results.

Case 3. RoSAS is also applicable to identify novel anomalies that do not appear during training, as validated in Case 3. This advantage owes to the feature learning module of RoSAS. The learning objective posed upon the representation space judges whether labeled anomalies are effectively separated by introducing a reference divergence degree between unlabeled data. That is, this learning objective not only repels anomalies from the normal manifold but pulls unlabeled samples together. Therefore, during inference, novel anomalies can be far away from the normal manifold in the representation space.

5.3. Robustness w.r.t. Anomaly Contamination Levels

This experiment evaluates the robustness of RoSAS w.r.t. different anomaly contamination ratios (i.e., the proportion of anomalies in the unlabeled set $\mathcal{X}_U$). As anomalies are rare events in practical scenarios, we vary the contamination level from 0% up to 8%, and all the contamination levels use 30 random true anomalies as labeled data.

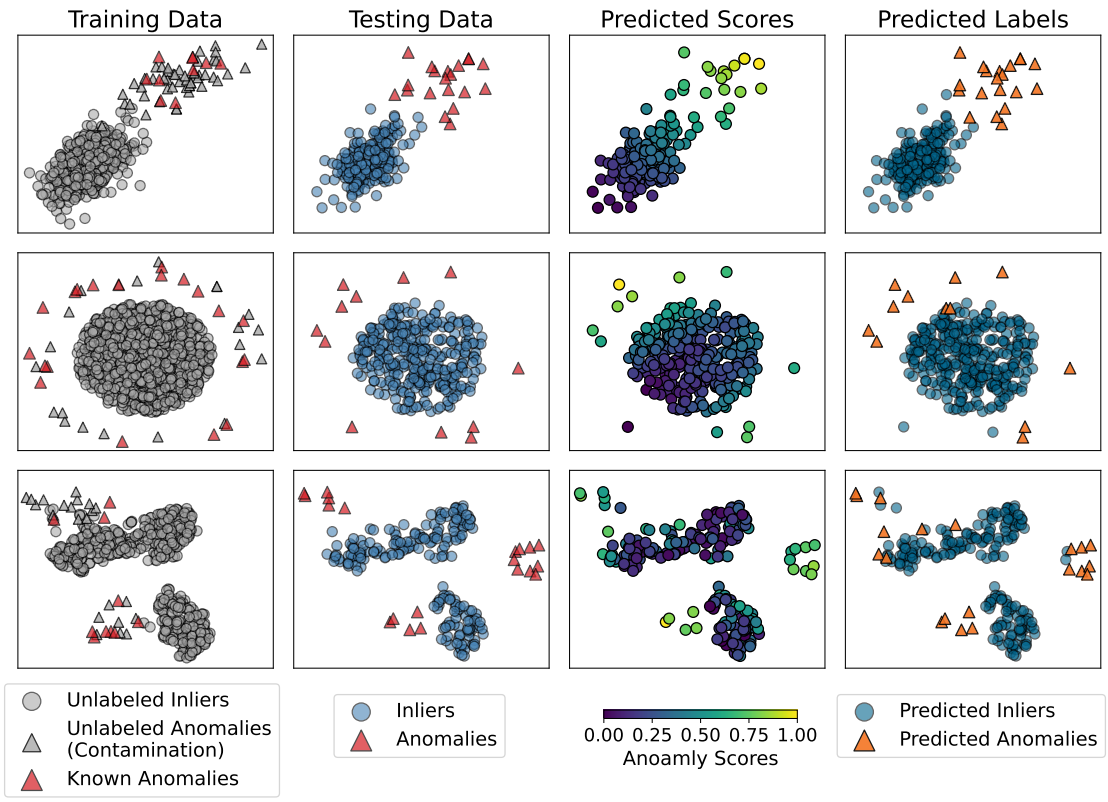


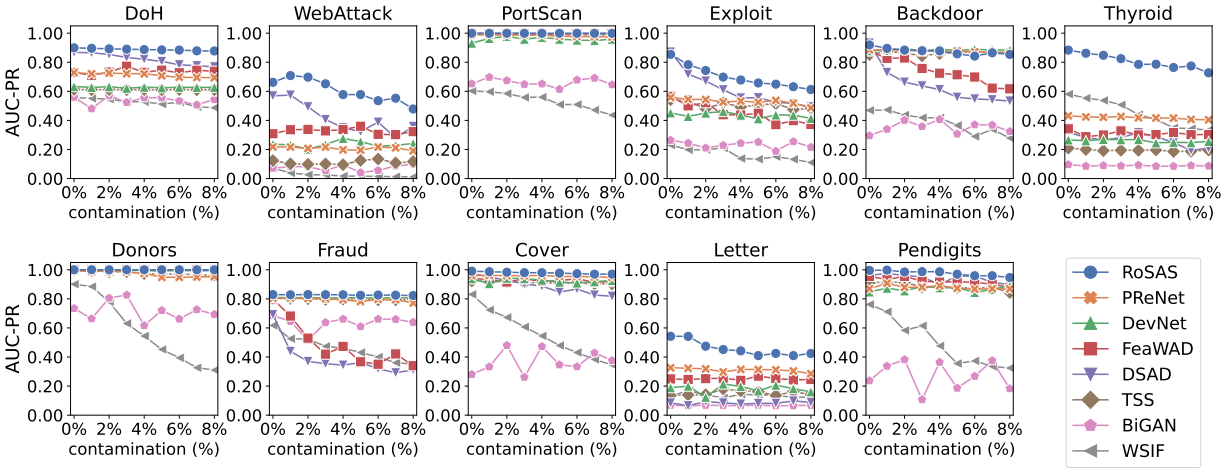
Figure 3: Three toy cases with different anomaly types. Each row indicates a case. The top and medium cases respectively contain *clustered anomalies* and *scattered anomalies*. The testing data of the bottom case has *novel anomalies* (the anomaly cluster on the far right) that do not appear in the training set. The panels in the left two columns show the data distribution of training/testing data, and the anomaly detection results of RoSAS including predicted anomaly scores and binary labels are visualized in the right two columns.

Figure 4 (a) shows the AUC-PR results on all the eleven real-world datasets with varying anomaly contamination levels. Anomaly detection performance generally decreases when the contamination level increases. Nevertheless, in the vast majority of cases, RoSAS is more robust than the competitors. It is noteworthy that, in some datasets (e.g., *DoH*, *PortScan*, and *Pendigits*), most anomaly detectors are stable when increasing the contamination rate. It might be because the increased anomalies are isolated data samples, and anomaly detection models can easily filter the interference. The competitors can also obtain very competitive performance on these datasets. However, in complicated datasets like *Exploit*, *Backdoor*, *Thyroid*, and *Fraud*, these anomalies that are hidden in the unlabeled set greatly blur the boundary between normal data and anomalies. RoSAS can consistently obtain better performance than the competitors in challenging noisy environments with high contamination levels.

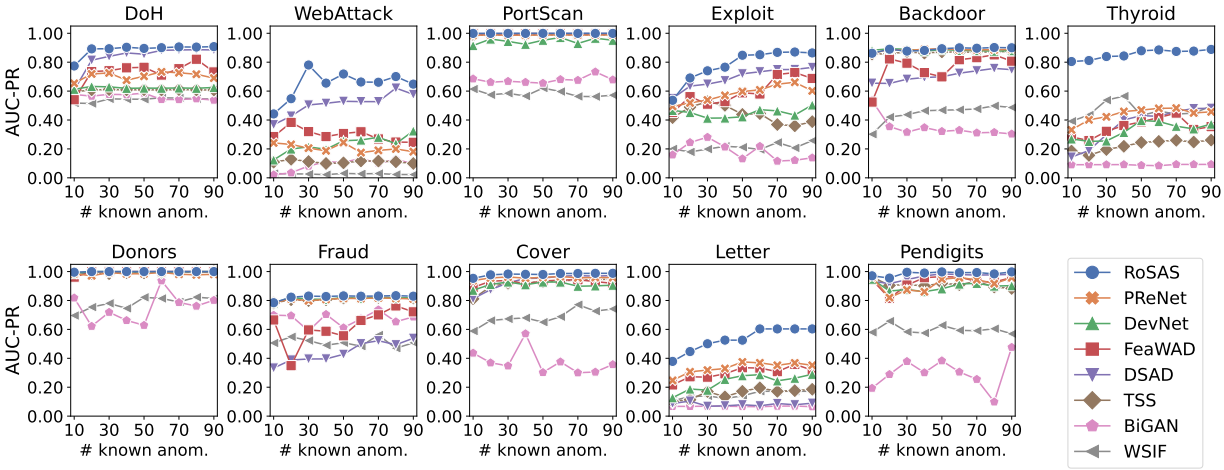
5.4. Data Efficacy of Labeled Anomalies

This experiment estimates the data efficacy of different numbers of labeled anomalies in terms of the value they bring to semi-supervised anomaly detection. In other words, this experiment examines whether RoSAS achieves more significant performance improvement than its competing methods when more labeled anomalies are available. The number of labeled anomalies is increased from 10 to 90, and the contamination level is maintained at 2%.

Figure 4 (b) shows the AUC-PR results of RoSAS and its contenders w.r.t varying numbers of labeled anomalies. Semi-supervised anomaly detectors generally perform better when more labeled anomalies are accessible. However, this law is not always true in practice. Some anomaly detectors also present fluctuation trends on some datasets. These increased labeled anomalies may have heterogeneous behaviors and carry conflicting information which imposes negative effects on anomaly detectors, as has been explained in (Pang et al., 2019). By contrast, our method obtains



(a)



(b)

Figure 4: AUC-PR results of RoSAS and its semi-supervised competing methods on datasets with **(a)** different anomaly contamination levels (i.e., ratios of anomalies in the unlabeled set $\mathcal{X}_U$) and **(b)** varying numbers of labeled anomalies (i.e., the size of labeled anomaly examples $\mathcal{X}_A$).

more stable and superior performance by fully utilizing limited labeled anomalies. It is noteworthy that some detectors also do not perform better when more labeled anomalies are available. It might be because these increased labeled anomalies have very similar behaviors and fail to bring useful information related to the anomaly distribution.

5.5. Scalability Test

This experiment evaluates the scalability of RoSAS. Nine datasets are created with the same data size (i.e., 5,000) and dimensionality increasing in multiples of 2 from 16 to 4,096. Another nine datasets are generated with varying data sizes increasing from 4,000 to 1,024,000 with fixed dimensions (i.e., 128). For the sake of comparison fairness, we employ deep semi-supervised anomaly detection methods (i.e., PReNet, DevNet, FeaWAD, DSAD, TSS, and BiGAN) as counterparts in this experiment. We use the same training configuration for these methods including the size of mini-batches (32) and the number of mini-batches per training epoch (20). We report the execution time including the training time of 10 epochs and the inference time. Scalability test results are reported in Figure 5. RoSAS and its counterparts can efficiently handle high-dimensional data thanks to the parallel accelerators in the mini-batch calculation of GPU. RoSAS only takes less than 10 seconds when handling 4,096-dimensional data. In terms of the scale-up test w.r.t. data size, RoSAS, DSAD, and TSS have comparably good efficiency. In comparison,

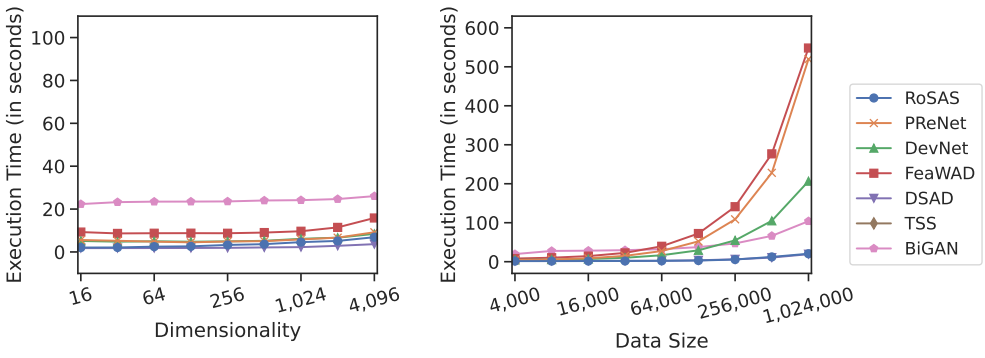


Figure 5: Scalability test results.

Table 4: AUC-PR results of RoSAS and its five ablated versions with the improvement rates of RoSAS compared to its variants per dataset. Positive rates are boldfaced.

Data	**RoSAS**	$L \rightarrow L_{dis}$	$L \rightarrow L_{dev}$	$L \rightarrow L_{reg}$	**w/o** L'	**w/o** L_c
DoH	$0.893_{\pm 0.005}$	$0.894_{\pm 0.006}$ (-0.1%)	$0.888_{\pm 0.011}$ (**0.6%**)	$0.867_{\pm 0.008}$ (**3.0%**)	$0.892_{\pm 0.006}$ (**0.1%**)	$0.891_{\pm 0.005}$ (**0.2%**)
WebAttack	$0.781_{\pm 0.051}$	$0.623_{\pm 0.146}$ (**25.4%**)	$0.508_{\pm 0.136}$ (**53.7%**)	$0.434_{\pm 0.090}$ (**80.0%**)	$0.542_{\pm 0.081}$ (**44.1%**)	$0.638_{\pm 0.139}$ (**22.4%**)
PortScan	$0.999_{\pm 0.000}$	$0.999_{\pm 0.000}$ (0.0%)	$0.999_{\pm 0.000}$ (0.0%)	$0.997_{\pm 0.001}$ (**0.2%**)	$0.997_{\pm 0.001}$ (**0.2%**)	$0.999_{\pm 0.000}$ (0.0%)
Exploit	$0.740_{\pm 0.026}$	$0.727_{\pm 0.031}$ (**1.8%**)	$0.674_{\pm 0.038}$ (**9.8%**)	$0.632_{\pm 0.069}$ (**17.1%**)	$0.694_{\pm 0.031}$ (**6.6%**)	$0.721_{\pm 0.034}$ (**2.6%**)
Backdoor	$0.877_{\pm 0.021}$	$0.873_{\pm 0.020}$ (**0.5%**)	$0.858_{\pm 0.024}$ (**2.2%**)	$0.888_{\pm 0.011}$ (-1.2%)	$0.892_{\pm 0.009}$ (-1.7%)	$0.877_{\pm 0.020}$ (0.0%)
Thyroid	$0.839_{\pm 0.046}$	$0.623_{\pm 0.167}$ (**34.7%**)	$0.543_{\pm 0.190}$ (**54.5%**)	$0.777_{\pm 0.021}$ (**8.0%**)	$0.812_{\pm 0.049}$ (**3.3%**)	$0.833_{\pm 0.044}$ (**0.7%**)
Donars	$1.000_{\pm 0.000}$	$1.000_{\pm 0.000}$ (0.0%)	$1.000_{\pm 0.000}$ (0.0%)	$1.000_{\pm 0.000}$ (0.0%)	$1.000_{\pm 0.000}$ (0.0%)	$1.000_{\pm 0.000}$ (0.0%)
Fraud	$0.831_{\pm 0.003}$	$0.830_{\pm 0.003}$ (**0.1%**)	$0.824_{\pm 0.008}$ (**0.8%**)	$0.820_{\pm 0.005}$ (**1.3%**)	$0.819_{\pm 0.005}$ (**1.5%**)	$0.830_{\pm 0.004}$ (**0.1%**)
Cover	$0.983_{\pm 0.003}$	$0.980_{\pm 0.005}$ (**0.3%**)	$0.982_{\pm 0.004}$ (**0.1%**)	$0.973_{\pm 0.005}$ (**1.0%**)	$0.985_{\pm 0.002}$ (-0.2%)	$0.983_{\pm 0.004}$ (0.0%)
Letter	$0.501_{\pm 0.026}$	$0.433_{\pm 0.059}$ (**15.7%**)	$0.417_{\pm 0.039}$ (**20.1%**)	$0.333_{\pm 0.049}$ (**50.5%**)	$0.372_{\pm 0.050}$ (**34.7%**)	$0.494_{\pm 0.064}$ (**1.4%**)
Pendigits	$0.995_{\pm 0.013}$	$0.990_{\pm 0.014}$ (**0.5%**)	$0.987_{\pm 0.015}$ (**0.8%**)	$0.978_{\pm 0.011}$ (**1.7%**)	$0.999_{\pm 0.001}$ (-0.4%)	$0.994_{\pm 0.012}$ (**0.1%**)
Average	$0.858_{\pm 0.018}$	$0.816_{\pm 0.041}$ (**7.1%**)	$0.789_{\pm 0.042}$ (**13.0%**)	$0.791_{\pm 0.025}$ (**14.7%**)	$0.819_{\pm 0.021}$ (**8.0%**)	$0.842_{\pm 0.030}$ (**2.5%**)
p-value	-	0.013	0.008	0.014	0.126	0.018

FeaWAD and PReNet have complicated network structures that lead to significantly increased execution time. RoSAS uses about 20 seconds to handle the dataset containing 1,024,000 data samples.

5.6. Ablation Study

This experiment is to validate the contribution of key designs in RoSAS. We set five ablated versions. The changes in these variants are introduced as follows, and other parts are the same as RoSAS.

- $\boldsymbol{L \rightarrow L_{dis}}$ discretizes the generated continuous supervision targets used in the anomaly scoring loss function L.
- $\boldsymbol{L \rightarrow L_{dev}}$ uses state-of-the-art anomaly scoring loss function used in DevNet (Pang et al., 2019, 2021a) to replace L.
- $\boldsymbol{L \rightarrow L_{reg}}$ uses a bare regression loss function used in RoSAS (i.e., smooth-ℓ_1 loss) to replace L.
- **w/o** $\boldsymbol{L'}$ removes the feature learning-based regularizer L'.
- **w/o** $\boldsymbol{L_c}$ removes the consistency learning part in L.

The first three variants ($\boldsymbol{L \rightarrow L_{dis}}$, $\boldsymbol{L \rightarrow L_{dev}}$, and $\boldsymbol{L \rightarrow L_{reg}}$) only take *discrete supervision information* to optimize anomaly scores, which are used to verify the significance of continuous supervision-guided anomaly score optimization. The ablated variants **w/o** $\boldsymbol{L'}$ and **w/o** $\boldsymbol{L_c}$ measure the contributions of the feature learning-based regularization L' and the consistency constraint in L.

The AUC-PR performance of RoSAS and its five ablated variants is shown in Table 4. RoSAS significantly outperforms its three ablated versions $\boldsymbol{L \rightarrow L_{dis}}$, $\boldsymbol{L \rightarrow L_{dev}}$, and $\boldsymbol{L \rightarrow L_{reg}}$ at 98% confidence level. More than 7% average improvement rate is achieved. These three variants use various objectives with only discrete supervision information.

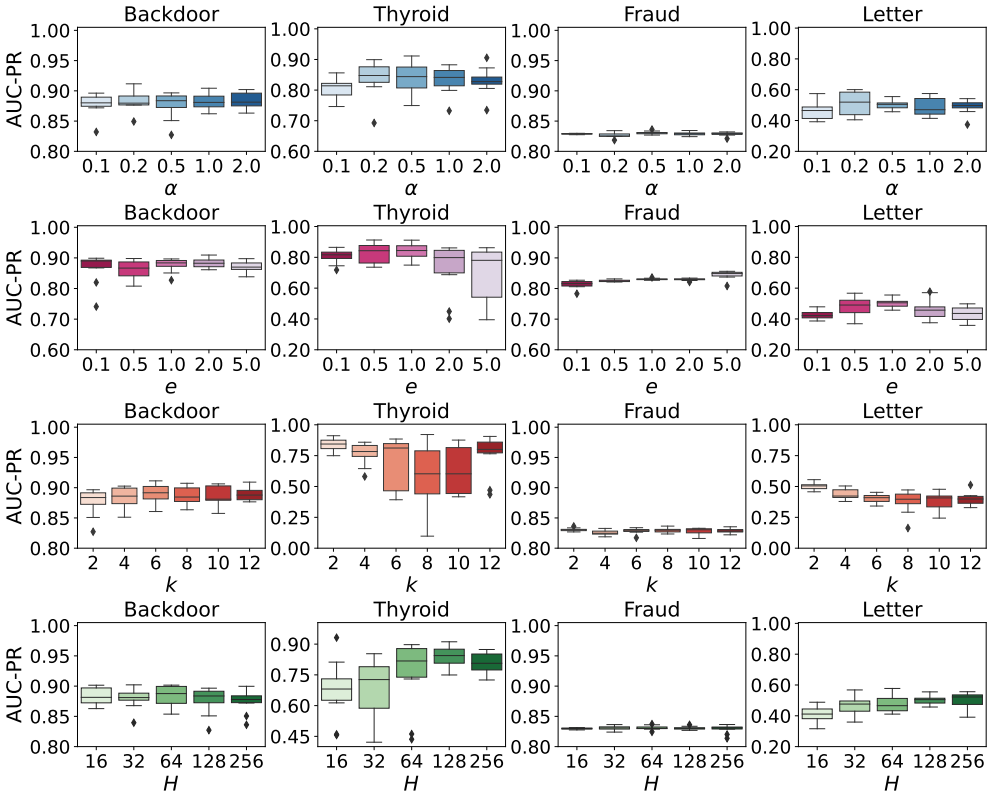


Figure 6: AUC-PR results of RoSAS with different settings of four key hyper-parameters (α, e, k, and H).

This comparison result validates the significance of using continuous supervision signals in anomaly score optimization. Anomalies are with various abnormal degrees, and anomaly scores naturally take on a continuous distribution. It is hard for discrete supervision information to accurately describe such consecutive trends in continuous distribution, resulting in suboptimal optimization of anomaly scores in these real-world datasets. Our work reveals this significant limitation in current anomaly detection studies and devises a simple but effective solution. On the other hand, RoSAS outperforms **w/o** $\boldsymbol{L'}$ by 8% and at 85% confidence level, which verifies the complementary robustness enhancement effect brought by the feature learning-based regularization. Compared to **w/o** $\boldsymbol{L_c}$, the average improvement is above 2% and the confidence interval is 98%, which quantitatively measures the contribution of the consistency learning in producing smoother anomaly scores.

5.7. Sensitivity Test

We investigate the influence of different settings of key hyper-parameters in RoSAS, i.e., α in the Beta distribution, k in the mass interpolation, e in the regularizer, and the intermediate representation dimension H. These hyper-parameters are tuned in turn and other parameters are kept the same as previously reported. RoSAS is performed 10 times on each hyper-parameter setting. The box plot of 10 AUC-PR values per dataset is illustrated in Figure 6. We show four representative datasets, and the other seven datasets are with similar or stable trends. As analyzed before, α may influence the detection performance to some extent, and we use $\alpha = 0.5$ considering the "manifold intrusion" problem. Besides, we can safely use a margin $e = 1$ in the feature learning-based regularizer. The choice of k might considerably influence the detection performance, and $k = 2$ is more stable. Lower representation dimension H fails to convey sufficient information to the downstream anomaly scoring process, and thus 128 is recommended.

6. Discussion and Implications

6.1. Key contributions

This study first summarizes the prior arts of this research field by giving a general semi-supervised anomaly detection framework. This framework contributes a unifying view of this research line, and we theoretically show how representative existing methods are instantiated from this framework. It may also offer valuable insights into the design of new semi-supervised anomaly detection models. More importantly, we uncover the key limitations of supervisory signals directly supplied by the semi-supervised setting and broadly used in existing methods. Motivated by these problems, we further propose a concrete anomaly detection method, and specifically, we make the following technical contributions.

This study contributes to the semi-supervised anomaly detection literature by taking into account the anomaly contamination problem. Arguably, many prior works directly use the whole unlabeled set as normal data for training their models (see Table 1 and Section 3.3), and their performance is considerably downgraded by these noisy points (as illustrated in the toy case in Figure 1 and real-world datasets in Table 3). Our method RoSAS is shown to be a simple yet effective solution to address this limitation. Instead of directly feeding original flawed supervision into the learning model, we propose new supervision containing augmented data with more reliable label information, resulting in stronger robustness than existing state-of-the-art methods when the training set is with high contamination level (see empirical results in Figure 4).

We consider our work as a starting point for leveraging continuous supervision information to optimize continuously distributed anomaly scores. To the best of our knowledge, we are the first to raise this issue in anomaly detection. We empirically show the advantage of using continuous supervision over discretized ones (see Table 4). Continuous supervision can lead to significant performance gain at 98% confidence level. Also, we pose a consistency constraint to further enhance the capability of producing smoother anomaly scores, which brings about 5% performance improvement. These findings may foster future theoretical research or inspire new optimization mechanisms of anomaly scores.

To sum up, different from current studies that rely on *contaminated discrete supervision*, our core novelty is a new kind of *contamination-resilient continuous supervision*. This supervision better conforms to the real abnormal-normal distribution and offers significantly better guidance to the optimization of the end-to-end anomaly scoring neural network.

6.2. Practical Implications

Albeit a plethora of unsupervised anomaly detection models, many real-world systems are looking for anomaly detectors that can exploit their historical anomalies, and this study adds a new competitive option to the list that currently only contains limited choices. We show that only 30 anomalies can bring drastically improved performance than unsupervised models that work on unlabeled data only (e.g., our approach RoSAS achieves 0.999 of AUC-PR on an intrusion detection dataset *PortScan* while unsupervised performance is only as low as 0.1). Given such huge benefits, instead of digging into the design of unsupervised anomaly detection models, one quick way to boost detection performance might be to transfer the unsupervised setting to the semi-supervised paradigm by feeding a few anomaly examples.

There are also many research and development fronts that we are pursuing in the future to further enhance the practical impact of this research. On one hand, this study can be extended to applications in different fields. We employ eleven datasets mainly from three domains including cybersecurity, medicine, and finance, and our approach also has the potential to identify system faults in AIOps or attacks in AI safety. On the other hand, by plugging in advanced network structures, our approach can be also applied to handle different data types (e.g., Transformer for sequential data, graph neural networks for graph data, and convolutional networks for images).

7. Conclusions

This paper first presents a general framework of deep semi-supervised anomaly detection to summarize this research line and reveal two key limitations of current studies. We then propose RoSAS, a concrete deep semi-supervised anomaly detection method. By optimizing the detection model using the mass-interpolation-based continuous supervision that explicitly indicates faithful abnormal degrees, RoSAS learns accurate and noise-tolerate anomaly scores.

Through extensive empirical results, we show two key advantages of using our continuous supervisory signals compared to the current discrete one: 1) our approach is substantially more robust w.r.t. anomaly contamination, especially on challenging cases with high contamination levels; 2) it is more data-efficient, that is, different numbers of labeled anomalies can be fully leveraged. These advantages are the main drivers of the overall superior performance of RoSAS that achieves about 20%-30% AUC-PR improvement over state-of-the-art semi-supervised anomaly detection approaches on 11 real-world datasets.

Acknowledgments

Hongzuo Xu, Yijie Wang, Songlei Jian, Ning Liu, and Yongjun Wang are supported in part by the National Key R&D Program of China under Grant 2022ZD0115302, in part by the National Natural Science Foundation of China under Grants 62002371 and 61379052, in part by the Science Foundation of Ministry of Education of China under Grant 2018A02002, in part by the Postgraduate Scientific Research Innovation Project of Hunan Province under Grants CX20210049 and CX20210028, in part by the Natural Science Foundation for Distinguished Young Scholars of Hunan Province under Grant 14JJ1026, and the Foundation of National University of Defense Technology under Grant ZK21-17. Guansong Pang is supported in part by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 under Grant 21SISSMU031.

We also thank the referees for their comments, which helped improve this paper considerably.

References

Bandaragoda, T. R., Ting, K. M., Albrecht, D., Liu, F. T., Zhu, Y., & Wells, J. R. (2018). Isolation-based anomaly detection using nearest-neighbor ensembles. *Computational Intelligence*, *34*, 968–998.

Barbariol, T., & Susto, G. A. (2022). Tiws-iforest: Isolation forest in weakly supervised and tiny ml scenarios. *Information Sciences*, *610*, 126–143.

Carmona, C. U., Aubet, F.-X., Flunkert, V., & Gasthaus, J. (2022). Neural contextual anomaly detection for time series. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence* (pp. 2843–2851).

Chen, K., Yao, L., Zhang, D., Wang, X., Chang, X., & Nie, F. (2019). A semisupervised recurrent convolutional attention model for human activity recognition. *IEEE Transactions on Neural Networks and Learning Systems*, *31*, 1747–1756.

Ding, C., Pang, G., & Shen, C. (2022). Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7388–7398).

Ding, K., Li, J., Bhanushali, R., & Liu, H. (2019). Deep anomaly detection on attributed networks. In *Proceedings of the 2019 SIAM International Conference on Data Mining* (pp. 594–602). SIAM.

Ding, K., Zhou, Q., Tong, H., & Liu, H. (2021). Few-shot network anomaly detection via cross-network meta-learning. In *Proceedings of the Web Conference* (pp. 2448–2456).

Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 315–324).

Golan, I., & El-Yaniv, R. (2018). Deep anomaly detection using geometric transformations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 9758–9769).

Gong, D., Liu, L., Le, V., Saha, B., Mansour, M. R., Venkatesh, S., & Hengel, A. v. d. (2019). Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1705–1714).

Guo, H., Mao, Y., & Zhang, R. (2019). Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 3714–3722). volume 33.

Han, S., Hu, X., Huang, H., Jiang, M., & Zhao, Y. (2022). Adbench: Anomaly detection benchmark. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track*.

Huang, C., Ye, F., Zhao, P., Zhang, Y., Wang, Y.-F., & Tian, Q. (2020). Esad: End-to-end deep semi-supervised anomaly detection. *arXiv preprint arXiv:2012.04905*, .

Huang, T., Chen, P., & Li, R. (2022). A semi-supervised vae based active anomaly detection framework in multivariate time series for online systems. In *Proceedings of the ACM Web Conference 2022* (pp. 1797–1806).

Jiang, M., Hou, C., Zheng, A., Hu, X., Han, S., Huang, H., He, X., Yu, P. S., & Zhao, Y. (2023). Weakly supervised anomaly detection: A survey. *arXiv preprint arXiv:2302.04549*, .

Kang, L., Liu, J., Liu, L., Zhou, Z., & Ye, D. (2021). Semi-supervised emotion recognition in textual conversation via a context-augmented auxiliary training task. *Information Processing & Management*, *58*, 102717.

Li, Z., Zhao, Y., Botta, N., Ionescu, C., & Hu, X. (2020). COPOD: copula-based outlier detection. In *Proceedings of the 20th IEEE International Conference on Data Mining* (pp. 1118–1123). IEEE.

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the 8th IEEE International Conference on Data Mining* (pp. 413–422). IEEE.

Liu, S., Johns, E., & Davison, A. J. (2019). End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1871–1880).

Lv, J., Wang, Y., & Chen, S. (2023). Adaptive multivariate time-series anomaly detection. *Information Processing & Management*, *60*, 103383.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, *9*.

Pang, G., Cao, L., Chen, L., & Liu, H. (2018). Learning representations of ultrahigh-dimensional data for random distance-based outlier detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2041–2050).

Pang, G., Ding, C., Shen, C., & Hengel, A. v. d. (2021a). Explainable deep few-shot anomaly detection with deviation networks. *arXiv preprint arXiv:2108.00462*, .

Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021b). Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, *54*, 1–38.

Pang, G., Shen, C., & van den Hengel, A. (2019). Deep anomaly detection with deviation networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 353–362).

Pang, G., Shen, C., Jin, H., & Hengel, A. v. d. (2023). Deep weakly-supervised anomaly detection. In *Proceedings of the 29th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

Pang, G., Yan, C., Shen, C., Hengel, A. v. d., & Bai, X. (2020). Self-trained deep ordinal regression for end-to-end video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12173–12182).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ruff, L., Vandermeulen, R., Goernitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., & Kloft, M. (2018). Deep one-class classification. In *Proceedings of the International Conference on Machine Learning* (pp. 4393–4402).

Ruff, L., Vandermeulen, R. A., Görnitz, N., Binder, A., Müller, E., Müller, K.-R., & Kloft, M. (2020). Deep semi-supervised anomaly detection. In *International Conference on Learning Representations*.

Shenkar, T., & Wolf, L. (2022). Anomaly detection for tabular data with internal contrastive learning. In *International Conference on Learning Representations*.

Tax, D. M., & Duin, R. P. (2004). Support vector data description. *Machine learning*, *54*, 45–66.

Tian, B., Su, Q., & Yin, J. (2022). Anomaly detection by leveraging incomplete anomalous knowledge with anomaly-aware bidirectional gans. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence* (pp. 2255–2261).

Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine learning*, *109*, 373–440.

Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., & Van Gool, L. (2021). Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*, 3614–3633.

Wu, Z., Xu, H., Wang, Y., & Wang, Y. (2021). Surrogate supervision-based deep weakly-supervised anomaly detection. In *Proceedings of the 21st IEEE International Conference on Data Mining Workshops* (pp. 975–982). IEEE.

Xu, H., Pang, G., Wang, Y., & Wang, Y. (2023a). Deep isolation forest for anomaly detection. *IEEE Transactions on Knowledge and Data Engineering*, (pp. 1–14). doi:10.1109/TKDE.2023.3270293.

Xu, H., Wang, Y., Jian, S., Huang, Z., Wang, Y., Liu, N., & Li, F. (2021a). Beyond outlier detection: Outlier interpretation by attention-guided triplet deviation network. In *Proceedings of the Web Conference* (pp. 1328–1339).

Xu, H., Wang, Y., Wang, Y., & Wu, Z. (2019). Mix: A joint learning framework for detecting both clustered and scattered outliers in mixed-type data. In *Proceedings of the 19th IEEE International Conference on Data Mining* (pp. 1408–1413). IEEE.

Xu, H., Wang, Y., Wei, J., Jian, S., Li, Y., & Liu, N. (2023b). Fascinating supervisory signals and where to find them: Deep anomaly detection with scale learning. In *Proceedings of the International Conference on Machine Learning*.

Xu, Y.-X., Pang, M., Feng, J., Ting, K. M., Jiang, Y., & Zhou, Z.-H. (2021b). Reconstruction-based anomaly detection with completely random forest. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)* (pp. 127–135). SIAM.

Yu, E., Sun, J., Li, J., Chang, X., Han, X.-H., & Hauptmann, A. G. (2018). Adaptive semi-supervised feature selection for cross-modal retrieval. *IEEE Transactions on Multimedia*, *21*, 1276–1288.

Zhang, C., Song, D., Chen, Y., Feng, X., Lumezanu, C., Cheng, W., Ni, J., Zong, B., Chen, H., & Chawla, N. V. (2019). A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the AAAI Conference on Artificial Intelligence* (pp. 1409–1416).

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018a). mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*.

Zhang, Y.-L., Li, L., Zhou, J., Li, X., Liu, Y., Zhang, Y., & Zhou, Z.-H. (2017). Poster: A pu learning based system for potential malicious url detection. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security* (pp. 2599–2601).

Zhang, Y.-L., Li, L., Zhou, J., Li, X., & Zhou, Z.-H. (2018b). Anomaly detection with partially observed anomalies. In *Companion Proceedings of the Web Conference* (pp. 639–646).

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *Journal of Machine Learning Research*, *20*, 1–7.

Zhou, S., Huang, X., Liu, N., Tan, Q., & Chung, F.-L. (2022). Unseen anomaly detection on networks via multi-hypersphere learning. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)* (pp. 262–270). SIAM.

Zhou, Y., Song, X., Zhang, Y., Liu, F., Zhu, C., & Liu, L. (2021). Feature encoding with autoencoders for weakly supervised anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, *33*, 2454–2465.