Dai, Yongsheng; Wang, Hui; Rafferty, Karen; Spence, Ivor; Quinn, Barry

Working Paper

TDSRL: Time Series Dual Self-Supervised Representation Learning for Anomaly Detection from Different Perspectives

QBS Research Paper, No. 2024/03

Provided in Cooperation with:
Queen's University Belfast, Queen's Business School

Suggested Citation: Dai, Yongsheng; Wang, Hui; Rafferty, Karen; Spence, Ivor; Quinn, Barry (2024) : TDSRL: Time Series Dual Self-Supervised Representation Learning for Anomaly Detection from Different Perspectives, QBS Research Paper, No. 2024/03, Queen's University Belfast, Queen's Business School, Belfast

This Version is available at:
https://hdl.handle.net/10419/289582

QUEEN'S UNIVERSITY BELFAST
QUEEN'S BUSINESS SCHOOL

Working Paper Series - QBS Research Paper 2024/03

TDSRL: Time Series Dual Self-Supervised Representation Learning for Anomaly Detection from Different Perspectives

Yongsheng Dai
Queen's University Belfast

Hui Wang
Queen's University Belfast

Karen Rafferty
Queen's University Belfast

Ivor Spence
Queen's University Belfast

Barry Quinn
Queen's University Belfast

10 April 2024

Series edited by Philip T. Fliers and Louise Moss.
To subscribe click here.
To submit forward your paper to qbs.rps@qub.ac.uk.

TDSRL: Time Series Dual Self-Supervised Representation Learning for Anomaly Detection from Different Perspectives

Yongsheng Dai, Hui Wang, *Member, IEEE*, Karen Rafferty, *Member, IEEE*,
Ivor Spence, *Member, IEEE*, Barry Quinn, *Member, IEEE*

***Abstract*—Time series anomaly detection plays a critical role in various applications, from finance to industrial monitoring. Effective models need to capture both the inherent characteristics of time series data and the unique patterns associated with anomalies. While traditional forecasting-based and reconstruction-based approaches have been successful, they tend to struggle with complex and evolving anomalies. For instance, stock market data exhibits complex and ever-changing fluctuation patterns that defy straightforward modelling. In this paper, we propose a novel approach called TDSRL (Time Series Dual Self-Supervised Representation Learning) for robust anomaly detection. TDSRL leverages synthetic anomaly segments which are artificially generated to simulate real-world anomalies. The key innovation lies in dual self-supervised pretext tasks: one task characterises anomalies in relation to the entire time series, while the other focuses on local anomaly boundaries. Additionally, we introduce a data degradation method that operates in both the time and frequency domains, creating a more natural simulation of real-world anomalies compared to purely synthetic data. Consequently, TDSRL is expected to achieve more accurate predictions of the location and extent of anomalous segments. Our experiments demonstrate that TDSRL outperforms state-of-the-art methods, making it a promising avenue for time series anomaly detection.**

***Index Terms*—Time series anomaly detection, self-supervised representation learning, contrastive learning, synthetic anomaly**

I. Introduction

With advancements in computational processes and sensor technology, time series data have become increasingly important in diverse applications such as IoT systems, clinical diagnosis, traffic analysis, financial supervision and climate science [1]–[4]. In analysing time series data, anomalies, or unusual patterns, have attracted significant attention from researchers due to their potential to indicate exceptional situations or events within a system, often with implications for safety or stakeholders' interests. Therefore, there is a growing demand for accurate time series anomaly detection.

Modern time series anomaly detection methods need to learn a representation model that allows effective encoding of the time series feature and identification of the unique pattern of anomalies. However, there is a fundamental challenge: the scarcity of anomalous data and the imbalance between normal and anomalous samples [5]. Anomalies are rare and may be obscured by normal data points, making data labelling difficult and expensive [6]. To address this challenge, unsupervised [7] or self-supervised [8] learning strategies have been widely adopted, as they do not require labelled data for training.

Unsupervised models, while flexible, have a number of limitations. Firstly, they often encounter subjective and environment-dependency problems. This is because we usually need to set various hyper parameters for such models, whose optimal values must be determined according to subjective experience and specific application contexts [9]. Secondly, the evaluation of unsupervised anomaly detection models may be a challenge for researchers [5]. Thirdly, and most critically, detecting anomalies from temporal contexts accurately without supervision remains a formidable task [6]. In particular, anomalies can exhibit unexpected behaviour and be related to multiple variables, making it difficult for unsupervised models to identify anomaly patterns based on the distribution of the data itself.

In contrast, self-supervised learning methods leverage pretext tasks to provide supervisory signals during training, leading to more robust anomaly detection models. However, constructing a self-supervised model places additional demands, including the need to design appropriate pseudo labels and pretext tasks [9], which can provide invaluable supervisory signals during the representation learning for unlabelled time series data, significantly improving model performance.

In this paper, we present a novel approach, the **T**ime Series **D**ual **S**elf-Supervised **R**epresentation **L**earning (TDSRL) network, for robust time series anomaly detection. TDSRL leverages dual self-supervised pretext tasks to model anomalies from both global and local perspectives. One task characterizes anomalies in relation to the entire time series, while the other focuses on local anomaly boundaries. Additionally, we propose a data degradation method that operates in both the time and frequency domains to generate synthetic anomalies, enhancing the model's ability to generalize across different anomaly types. Our proposed approach is concise and not dependent on any specific scenario. Thus, it offers a general paradigm for self-supervised time series representation modelling for anomaly detection tasks.

More specifically, we follow the idea of accomplishing self-supervised representation modelling based on synthetic anomalies [10], but we further extend and improve it. We first

Yongsheng Dai, Hui Wang, Karen Rafferty and Ivor Spence are with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, Northern Ireland (e-mail: ydai09@qub.ac.uk; h.wang@qub.ac.uk; K.Rafferty@qub.ac.uk; I.Spence@qub.ac.uk)

Barry Quinn is with Queen's Business School, Queen's University Belfast, Belfast, Northern Ireland (e-mail: b.quinn@qub.ac.uk)

introduce a data degradation method based on both time and frequency domains for anomaly synthesis, promoting a more natural simulation of real-world anomalies. The generation of synthetic anomalies can be achieved by adding a degradation or perturbation to the original time series. For this process, paper [10] and most previous work on time series anomaly detection focused only on the time domain. However, the frequency domain can provide insight into the behaviour of time series that cannot be captured solely in the time domain [1]. We believe that real-world time series anomalies should have significant changes in both time and frequency domains compared to normal intervals. We then transferred this concept to synthetic anomalies which serve as a foundation of creating pseudo labels and pretext tasks for self-supervised representation learning.

For pseudo labels, TDSRL employs a novel generation strategy based on equal-length segments of time series in a local domain. These new pseudo labels are positive-negative sample pairs composed of abnormal segments and adjacent normal segments. For the pretext task, TDSRL introduces a contrastive representation learning branch based on above pseudo labels. In the embedding space produced by TDSRL, this contrastive branch is aimed to enlarge the differences between the representation results of anomalous intervals and those of the segments before and after them. This contributes to the network having a higher discrimination ability between anomaly and normal within a local region in the downstream detection module. As a result, the network becomes more sensitive to both the start and end of anomalies, making it possible to detect the occurrence of anomalies in a more timely manner or even in advance. The overall success rate of anomaly detection will naturally increase accordingly.

In addition, the pretext task proposed in [10] primarily focuses on the relationship between anomalous segments and the entire time series on a global scale. We apply our synthetic anomalies to above 2 pretext tasks. Eventually, we are able to build an anomaly detection network with two self-supervised learning branches, which can simultaneously perform representation modelling for time series from both global and local perspectives.

The main contributions of this paper are threefold:

1) We introduce a data degradation approach based on both time and frequency domains to generate synthetic anomaly for self-supervised learning, enabling more natural simulation of real-world anomalies and enhancing model generalisation across different anomalies.
2) We propose a novel contrastive pretext task based on local sub-segments, improving the model's discrimination between anomaly and adjacent normal intervals, resulting in more accurate predictions of the location and extent of anomalous segments.
3) We introduce a self-supervised time series anomaly detection network with dual representation learning branches, capable of characterizing anomalies globally and enhancing sensitivity to local anomaly boundaries. When anomalies occur, our network can detect them more promptly or even in advance.

These contributions advance the state-of-the-art in time series anomaly detection and offer a promising avenue for future research in this field.

II. Related Work

A. Approaches for Time Series Anomaly Detection with Deep Learning Networks

Time series anomaly detection is a focal point in the field of big data research, with Deep Learning (DL) emerging as a powerful and popular technology for this task [11], [12]. In constructing DL anomaly detection networks for time series data, two main approaches have been explored: Forecasting-based and Reconstruction-based methods [5].

Forecasting-based approaches aim to learn a predictive model to fit the given time series data and predict future values. Anomalies are identified if the difference between the predicted and original inputs exceed a certain threshold [13]. On the other hand, Reconstruction-based models encode subsequences of normal training data in latent spaces and detect anomalies by reconstructing sliding windows from test data and comparing them to actual values, known as reconstruction error [14], [15].

In real-world scenarios, time series data may change rapidly or be unknown at any given moment. In such cases, Reconstruction-based networks may be more effective than Forecasting-based ones. However, both approaches rely on the assumption that the unlabelled training set contains no anomalies, which may not hold true for all anomaly detection scenarios.

The TDSRL approach, introduced in this paper, addresses these limitations by leveraging dual self-supervised representation learning of time series based on synthesized abnormal segments. This approach fully explores the relationship between anomaly intervals and other data points, as well as the unique characteristics of anomalies, without being constrained by specific scenarios, thus offering good generalizability.

B. Learning Schemes

Training deep learning networks for anomaly detection involves four learning schemes: supervised, semi-supervised, unsupervised, and self-supervised, depending on the availability of annotations during training.

Supervised methods aim to learn class boundaries based on all labels in the training set but are not applicable to time series anomaly detection due to the unknown or improperly labeled nature of anomalies [6]. Semi-supervised methods train models based on a context where the dataset consists only of normal points, detecting deviations from this distribution as anomalies. However, these methods have limitations, as discussed earlier.

Unsupervised learning is flexible but faces challenges in accurately detecting anomalies without supervision and may suffer from subject and environment-dependency and evaluation difficulties [5], [9], [10].

Self-supervised learning, like unsupervised learning, does not rely on original annotations and tends to have more

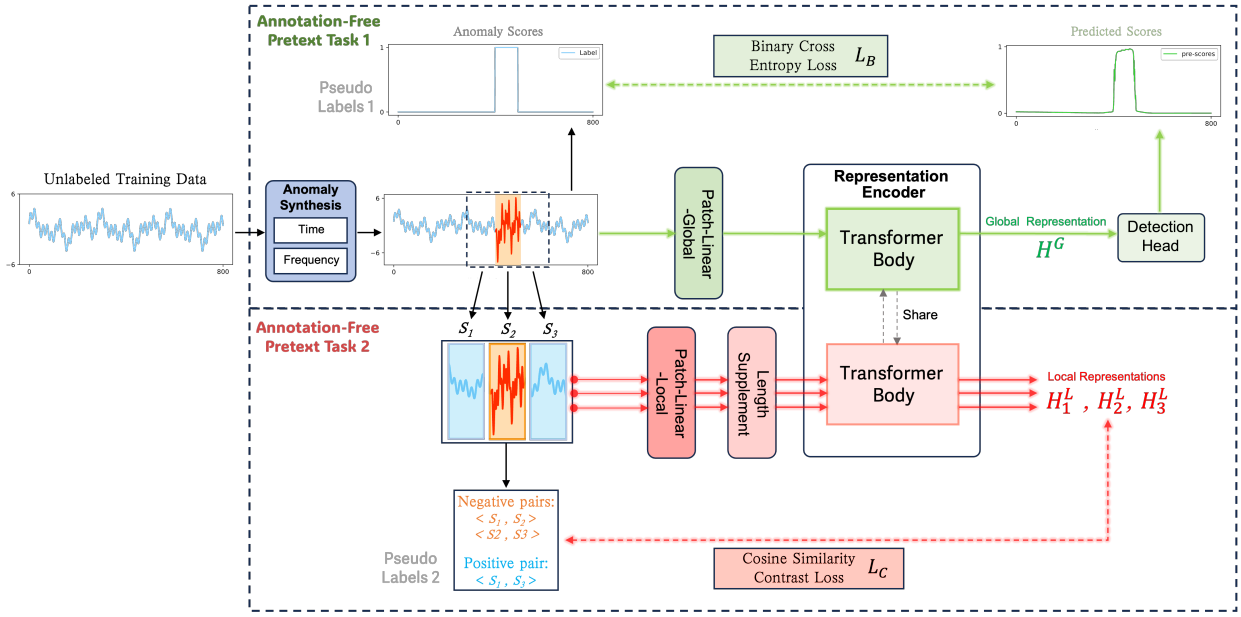

Fig. 1. The overall architecture of the proposed TDSRL. It mainly consists of 2 self-supervised representation learning branches. They have their own Pseudo Labels (1&2) and Pretext Tasks (1&2) derived from an Anomaly Synthesis module that considers both the time and frequency domains. The red and green lines represent the two representation learning processes. They jointly improve the performance of the Representation Encoder from a global and local perspective, which ultimately allows the detection head to detect anomalies in a more sensitive and timely manner.

stable performance. However, it requires designing appropriate pseudo labels and pretext tasks [16]. In this paper, we create these components based on synthetic anomaly segments for our dual self-supervised representation learning.

C. Contrastive Learning for Time Series

Contrastive learning is a widely used self-supervised learning strategy. The goal is to learn an encoder that maps inputs into an embedding space in which similar data samples (positive) are close to each other while dissimilar (negative) ones are far apart. It can help the network to distinguish any instance from the others more sensitively and accurately. Contrast learning was initially investigated mainly in computer vision tasks [17]. But it is also increasingly being applied to various time series problems. For instance, to enhance model generalization capacity to the target data in time series segmentation task, Xiao et al. [9] explored unlabelled target data using contrastive learning to enable the model to capture its characteristics. When building a time series pre-training model, Zhang et al. [1] simultaneously mined time domain information and frequency domain information based on contrastive learning. Eldele et al. [16] proposed a representation learning framework for time series classification task via temporal and contextual contrastive learning. Eldele et al. [18] employed a dual attention contrastive representation learning network in time series anomaly detection.

In these previous research works, people modeled the entire time series sample globally, and contrastive learning was only applied to two complete long sequences from different sources or processes. However, in the time series anomaly detection task, anomalies are sharp changes in data distribution, which are always rare and only occur in relatively very short segments. Applying the above methods directly to this task does not take advantage of contrastive learning to fully explore the huge differences between abnormal segments and their adjacent normal segments in the same long-term sequence. As a result, the network's sensitivity to the boundaries between these two types of fragments in the local area will not be sufficiently trained and improved. This will affect the network's ability to detect the occurrence of anomalies in a timely manner, which is critical for many application scenarios such as IoT and finance. This is also the main motivation for us to use contrastive learning to locally model the continuous sub-segments of the abnormal area in the same time series sample in the subsequent Pretext Task 2.

III. Method

A. Overview

Figure 1 shows the overall structure of our TDSRL. First, it possesses an anomaly synthesis module. It perturbs the original time series sub-segments based on both time-domain and frequency-domain Data Degradation strategies, thereby synthesizing artificial anomalies that are closer to real anomalies. They provide the foundation for the next step of self-supervised representation learning. The details of this module are elaborated in Section 3.2.

Second, Our time series representation model used for anomaly detection consists of two branches of Annotation-Free Pretext Tasks. They share the same representation encoder (the Transformer Body in Figure 1) and utilize their respective Pseudo Labels and self-supervised learning schemes to train the network simultaneously. Among them, Pretext Task 1 uses representation learning of the entire time series to mine the connection between abnormal and normal points from a global perspective. Meanwhile, Pretext Task 2 accomplishes self-supervised training based on contrastive learning, focusing

more on the local representation learning of abnormal areas. Its purpose is to further explore the differences and uniqueness of characteristics between anomaly segments and their adjacent normal segments.

B. Anomaly Synthesis with Time and Frequency-based Data Degradation

All time series are composed of both time-domain signals and frequency-domain signals, and the frequency information has been playing a key role in classic signal processing [19]. Thus we believe that real time series anomalies should exhibit anomalous changes or data degradation in both of these domains. However, most previous works (such as [10]) only consider adding perturbations to the original data in the time domain when synthesizing artificial anomalies for self-supervised training of anomaly detectors. This hinders the synthetic anomalies from fully reflecting the intrinsic characteristics of real anomalies.

In our method, we first randomly cut a sub-segment $X_s = X[t_a : t_a + l_a] \in \mathbb{R}^{l_a \times d}$ from a training sample $X \in \mathbb{R}^{T \times d}$ of the original time series, where T is the original sequence length, d is the number of input variables, l_a represents the size of synthetic anomaly window, the starting index t_a of X_s is randomly selected. Subsequently, we apply data degradation methods based on both time and frequency to X_s, thereby obtaining multiple types of synthesized abnormal segments $X_s^{'}$. The purpose is to mimic real-world anomalies more naturally and enhance the generalization ability of our self-supervised representation model to different anomalies.

For Frequency-based Data Degradation, we start by extracting the frequency spectrum of X_s using a transform operator such as Fourier Transformation [20]. Every frequency component in the frequency spectrum denotes a basis function of original time series with the corresponding frequency and amplitude. Here, we can divide the frequency spectrum into a high-frequency part (High-bank) and a low-frequency part (Low-bank). They respectively carry different types of information in time series. Our Frequency-based Data Degradation is mainly achieved by perturbing the frequency spectrum in these two ranges.

Specifically, we adopt the following four data degradation mechanisms in the frequency domain of segment X_s:

1) Low-bank perturbation: It is achieved by adding or removing frequency components in the low-frequency part of the frequency spectrum. When removing, we randomly select several frequency components and set their amplitudes to zero. When adding, we randomly select several positions in the frequency spectrum and add frequency components with $\alpha \cdot Am$. The Am is the maximum amplitude in the original frequency spectrum and α is a predefined range.
2) High-bank perturbation: Similar perturbations introduced in (1) are applied in the high-frequency part of the frequency spectrum.
3) Mixture-bank perturbation: Perturbations from (1) are applied in both low and high-frequency parts of the frequency spectrum.

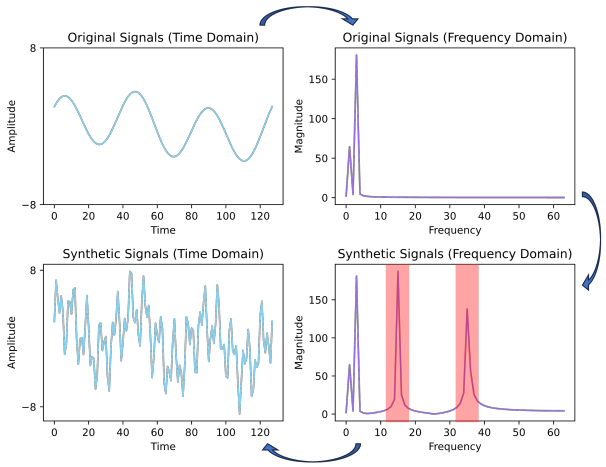

Fig. 2. An example of Frequency-based Data Degradation for synthetic anomaly segments generation.

4) Components Scaling: Randomly selecting several existing frequency components and scaling their values up or down within a predefined range.

An example of the above data degradation mechanism (3) is visualised in Figure 2. It can be seen that a small perturbation in the frequency domain may cause large changes to the temporal patterns in the time domain.

For Time-based Data Degradation, we adopted the four mechanisms proposed in [10]: Soft Replacement, Uniform Replacement, Peak Noise, and Length Adjustment. All of these mechanisms apply perturbations in the time domain of sub-segment X_s, that is, directly disturbing the values of the original time series. Ultimately, relying on both frequency domain and time domain data degradation mechanisms, we can obtain a total of eight types of artificially synthesized anomalies for subsequent self-supervised representation learning.

Finally, we replace the original segment in X with the 8 types of synthetic anomalies above, thus obtaining the new degraded samples $X^{'} \in \mathbb{R}^{T \times d}$ for subsequent dual self-supervised learning. Among them, the anomaly segments synthesized through frequency data degradation are converted into corresponding time-domain sequences before being embedded into the original time series. The overall process of generating $X^{'}$ is shown in Algorithm 1.

C. Pretext task 1: Global representation learning

To achieve accurate time series anomaly detection, it is essential to first construct a robust time series representation encoder (likes the Transformer Body shown in Figure 1), which is used to encode raw data into high-dimensional features. The process of detecting anomalies can be viewed as a downstream task based on these representation results. Therefore, whether the representation network and its encoder can effectively extract the intrinsic correlations and characteristics of the input time series (including abnormal segments) during the encoding process will have a significant impact on the accuracy of anomaly detection.

Considering the task context of lacking annotations and the inherent advantages of self-supervised learning, we will utilize

Algorithm 1 Generate Degraded Samples: $X^{'} \in \mathbb{R}^{T \times d}$

Input: Original time series samples $X \in \mathbb{R}^{T \times d}$, predefined proportion threshold Ψ, the size of synthetic anomaly window l_a

randomly select a starting index t_a in X
$X_s = X[t_a : t_a + l_a] \in \mathbb{R}^{l_a \times d}$
randomly initialize variable $\psi \in [0, 1]$
if $\psi < \Psi$ **then**
    # Time-based Degradation
    $X_s^{'} = g_{TD}(X_s)$
        ▷ g_{TD}: randomly apply 1 of the 4 types of Time-based Data Degradation
else
    # Frequency-based Degradation
    obtain frequency spectrum
    $X_s^f = g_{FT}(X_s)$ ▷ g_{FT}: Fourier Transform
    $X_s^{f'} = g_{FD}(X_s^f)$
        ▷ g_{FD}: randomly apply 1 of the 4 types of Frequency-based Data Degradation
    $X_s^{'} = g_{\hat{FT}}$ ▷ $g_{\hat{FT}}$: Inverse Fourier Transform
end if
$X^{'} = X.copy()$
$X^{'}[t_a : t_a + l_a] = X_s^{'}$
return $X^{'}$

this learning scheme to train TDSRL. To implement self-supervised learning, we need to design pseudo labels for the original unlabeled data and use them to provide supervisory signals for the pretext task during training. Continuously improving the feature extraction capability of our representation network TDSRL through self-supervised training in the pretext task is our primary goal. We use two Pretext Tasks (1&2) in this work, corresponding to the two representation learning branches of TDSRL.

For Pretext Task 1, first, we utilize the new degraded samples $X^{'} \in \mathbb{R}^{T \times d}$ to create the binary Pseudo Label $\widehat{Y} \in \{0, 1\}^N$. It represents the anomaly score of each time point in the sample. The value is 1 at the positions corresponding to the anomaly intervals, and 0 elsewhere. Second, we use a Patch-Linear Embedding module and a Transformer Body to model the entire time series, obtaining the Global Representation Results H^G. Third, the Detection Head outputs the predicted anomaly scores $Y \in [0, 1]^N$ based on H^G. Lastly, we complete the training in Pretext Task 1 by optimizing the below Binary Cross Entropy Loss L_B. The idea of Pretext Task 1 can be traced back to [10].

$$L_B = -\frac{1}{N} \sum_{i=1}^{N} \left[\widehat{Y}_i log(Y_i) + (1 - \widehat{Y}_i) log(1 - Y_i) \right], \quad (1)$$

Compared to conventional Linear Embedding before Transformer, a patch-wise linear projector [18] is more helpful for modeling the continuous temporal context of time series. In detail, the Patch-Linear module we employed first splices P neighbouring points along the original channel dimension d to create a patch. Thus, the input time series $X^{'} \in \mathbb{R}^{T \times d}$ are patched as $X^{'} \in \mathbb{R}^{N \times (P \times d)} = \mathbb{R}^{N \times d_p}$, where N is the number of patches. Then, a linear projection operation is applied in the patched channel dimension d_p, and the shape of output is $z^G \in \mathbb{R}^{N \times d_E}$. Finally, the dependencies among patches are modeled by Transformer Body to obtain H^G. This representation encoder is stacked by L identical layers, each of which mainly consists of a multi-headed self-attention (MHA) module followed by a multi-layer perceptron (MLP) block. We also adopt pre-norm residual connections in our Transformer Body, which can produce more stable gradients [16], [21]. In summary, the representation results are computed as:

$$\tilde{z}_l = MHA(LayerNorm(z_{l-1})) + z_{l-1}, 1 \leq l \leq L, \quad (2)$$

$$z_l = MHA(LayerNorm(\tilde{z}_l)) + \tilde{z}_l, 1 \leq l \leq L, \quad (3)$$

where $z_0 = z^G$ and $H^G = z_L \in \mathbb{R}^{N \times d_E}$ in Pretext Task 1.

Although this Pretext Task 1 can already explore the connection between anomalous and normal points from a global perspective through the representation learning of the entire time series. However, we believe that the representation learning of local regions of anomalies is equally important, especially the differences and uniqueness of characteristics between anomaly and adjacent normal intervals should be fully explored. Therefore, in the next section, we originally propose a new pretext task based on the contrastive learning for adjacent sub-segments.

D. Pretext task 2: Local representation learning

In real-world time series anomaly detection tasks, if a model can distinguish between anomalous intervals and their adjacent normal intervals with high accuracy, particularly being sensitive to the boundaries between these two types of intervals, it would significantly enhance the performance of anomaly detection. This capability signifies a substantial improvement of the model's accuracy in predicting the location and size of anomalous segments.

To achieve the above objectives, we construct the Pretext Task 2 based on contrastive learning, focusing on the local areas of synthetic anomaly. Contrastive learning is a popular self-supervised learning strategy. It starts by extracting samples with the same attributes from the original data and organizing them into positive pairs. Similarly, samples with the different attributes are organized into negative pairs. The process of labelling samples as $\langle positive, negative \rangle$ pairs provides pseudo labels for self-supervised learning. Following

TABLE I
TABLE 1: OVERALL RESULTS ON REAL-WORLD MULTIVARIATE DATASETS. THE P, R AND F1 ARE THE PRECISION, RECALL AND F1-SCORE. ALL RESULTS ARE IN %. THE BEST RESULTS ARE IN BOLD

Method	SMAP			SWaT			SMD		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Deep-SVDD	89.93	56.02	69.04	80.42	84.45	82.39	78.54	79.67	79.10
DAGMM	86.45	56.73	68.51	89.92	57.84	70.40	67.30	49.89	57.30
LSTM	89.41	78.13	83.39	86.15	83.27	84.69	78.55	85.28	81.78
CL-MPPCA	86.13	63.16	72.88	76.78	81.50	79.07	82.36	76.07	79.09
LSTM-VAE	92.20	67.75	78.10	76.00	89.50	82.20	75.76	90.08	82.30
BeatGAN	92.38	55.85	69.61	64.01	87.46	73.92	72.9	84.09	78.10
OmniAnomaly	92.49	81.99	86.92	81.42	84.30	82.83	83.68	86.82	85.22
ITAD	82.42	66.89	73.85	63.13	52.08	57.08	86.22	73.71	79.48
THOC	92.06	89.34	90.68	83.94	86.36	85.13	79.76	**90.95**	84.99
InterFusion	89.77	88.52	89.14	80.59	85.58	83.01	87.02	85.43	86.22
TS-CP2	87.65	83.18	85.36	81.23	74.10	77.50	87.42	66.25	75.38
SES-AD	89.35	78.73	83.70	90.98	85.53	88.17	34.66	78.49	48.09
TicTok	92.50	95.42	93.94	92.25	84.95	88.45	81.04	88.82	84.75
AnomalyBERT	94.47	88.67	91.48	93.09	90.64	91.85	93.24	74.21	82.64
TDSRL	**96.03**	**97.49**	**96.75**	**96.92**	**93.75**	**95.31**	**94.98**	87.70	**91.20**

this, the training goal of the contrastive pretext task is to urge the feature encoder to generate closer embeddings for positive pairs and to push the embeddings for negative pairs apart from each other. This will help downstream tasks to better identify and differentiate between these two types of samples.

In our task, as shown in the lower half of Figure 1, we individually segment the anomalous fragments and name them S_2. Simultaneously, we also segment the equally long normal fragments adjacent to S_2 and name them S_1 and S_3 respectively:

$$\begin{aligned} S_1 &= X^{'}[t_a - l_a - 1 : t_a - 1], \\ S_2 &= X^{'}[t_a : t_a + l_a], \\ S_3 &= X^{'}[t_a + l_a + 1 : t_a + 2 \times l_a + 1], \end{aligned} \tag{4}$$

Based on the discussion above, we use these three new samples to construct one positive pair $\langle S1, S3\rangle$ and two negative pairs $\langle S1, S2\rangle$ and $\langle S2, S3\rangle$. Subsequently, synchronising with Pretext Task 1 and sharing the same representation encoder (i.e., Transformer Body), Pretext Task 2 trains our TDSRL based on contrastive representation learning.

In detail, first, all these three time series segments are also projected into embedded features $z_i^L \in \mathbb{R}^{N^L \times d_E}$ with a Patch-Linear Embedding module, where $i = \{1, 2, 3\}$. Second, in order to align with the required shape of representation encoder, a Length Supplement module extends the the features $z_i^L \in \mathbb{R}^{N^L \times d_E}$ to $z_i^L \in \mathbb{R}^{N \times d_E}$ by the interpolation operation. Third, the shared Transformer Body utilizes z_i^L to calculate the Local Representation Results $H_i^L \in \mathbb{R}^{N \times d_E}$ of these three sub-segments. Finally, we use $D_{pos} = d(H_1^L, H_3^L)$, $D_{neg1} = d(H_1^L, H_2^L)$, $D_{neg2} = d(H_2^L, H_3^L)$ to denote the cosine distances of the representation results of $\langle positive, negative\rangle$ pairs, and input them into the contrastive loss function L_C to complete the training in Pretext Task 2. Inspired by the triplet loss [19], we design our Cosine Contrast Loss L_C as follows:

$$\begin{aligned} L_C = \sum_{D_{neg}} (D_{pos} - D_{neg} + \delta), \\ D_{neg} \in \{D_{neg1}, D_{neg2}\}, \end{aligned} \tag{5}$$

where δ is a predefined constant margin to ensure negative samples remain sufficiently distant with each other [22].

By continuously optimising the total loss L_{total} below, the training of Pretext Task 1 and Pretext Task 2 are performed simultaneously. Their main goal is to jointly learn a powerful representation encoder for time series anomaly detection.

$$L_{total} = L_B + L_C \tag{6}$$

E. Inference

The model's inference process will only apply Pretext Task 1, which can use the Detection Head to predict the anomaly score for each time point in the test set. An anomaly threshold will be set, and time points with anomaly scores exceeding this threshold are identified as anomalies. The specific value of the anomaly threshold is determined by validation set.

Although the inference process does not use Pretext Task 2, during the self-supervised dual-branch training above, Pretext Task 2 has further improved the ability of our representation encoder to clearly distinguish anomalous points from other

massive normal points in the input time series. The distinguishability of the representation results of the anomalous intervals and their adjacent normal intervals has also been further improved, which will greatly benefit the performance of the downstream anomaly detection head.

IV. Experiments

A. Benchmark Datasets

We adopt the following three widely-used benchmarks from real-world applications to evaluate TDSRL:

- Soil Moisture Active Passive (SMAP) dataset [23]]: is a dataset of soil samples and telemetry information used by the Mars rover. It is provided by NASA and contains 25 variables.
- Secure Water Treatment (SWaT) dataset [24]: is collected from a real-world water treatment plant with 7 days of normal and 4 days of abnormal operation. It is a 51-dimension sensor-based dataset.
- Server Machine Dataset (SMD) [25]: is a five-week long dataset. It is collected from the machines in a compute cluster, stacking accessed traces of their resource utilizations with 38 dimensions.

All of them are large multivariate time series datasets containing hundreds of thousands of timestamps. Each dataset consists of an unlabeled training set and a labeled test set.

B. Baselines and Evaluation Criteria

We compare our network with these 14 baselines proposed in recent years for comprehensive evaluations: Deep-SVDD (2018) [26], DAGMM (2018) [27], LSTM (2018) [28], LSTM-VAE (2018) [29], CL-MPPCA (2019) [30], BeatGAN (2019) [31], OmniAnomaly (2019) [32], ITAD (2020) [33], THOC (2020) [34], InterFusion (2021) [35], TS-CP2 (2021) [36], SES-AD (2022) [37], TicTok (2023) [38], AnomalyBERT (2023) [10].

In terms of evaluation criteria, Precision, Recall and F1-score are widely used to evaluate the performance of time series anomaly detection. They can be formulated as follows:

$$Precision = \frac{TP}{TP + FP}, \tag{7}$$

$$Recall = \frac{TP}{TP + FN}, \tag{8}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall}). \tag{9}$$

where the TP, FP, and FN respectively denote the number of positive samples (abnormal points) that are successfully predicted as positive, negative samples (normal points) that are incorrectly predicted as positive, and positive samples (abnormal points) that are incorrectly predicted as negative.

At the same time, an evaluation technique known as Point Adjustment (PA) [39] has become popular in this task over the last few years, and we also employ it in our work. Specifically, if any observation in the ground truth abnormal segment is correctly detected, all observations in the segment are considered to be correctly detected.

TABLE II
The results of Ablation studies. The 2 components are the Frequency-based Degradation in synthetic anomalies generation and the contrastive learning branch (i.e., Pretext task 2), respectively.

Components		P	R	F1
Frequency Degradation	Contrastive Branch			
✗	✗	93.09	90.64	91.85
✗	✓	95.08	93.04	94.05
✓	✗	93.84	90.98	92.39
✓	✓	**96.91**	**93.75**	**95.31**

C. Implementation Details

We summarize all the default hyper-parameters as follows in our implementation. Our TDSRL network contains six Transformer encoder layers. The dimension of the hidden state is 512, and the number of attention heads is 6. The window length l_{as} for the synthetic anomaly segment $S2$ and the proportion of Time-based Degradation are respectively set to 80 and 50% by default, but we will investigate the effect of different values of them on the network in the subsequent Parameter Sensitivity study. Besides, all the experiments are implemented in PyTorch [26] with one NVIDIA A100 32GB GPU. Adam [27] with default parameter is applied for optimization. We set the initial learning rate to 10^{-4} and the batch size to 16. We also employ the early-stop mechanism during training. Most of the other hyper-parameters are set with reference to [10].

D. Main Results

We first evaluate our TDSRL with fourteen competitive baselines on three real-world multivariate datasets as shown in Table 1. It can be seen that our proposed TDSRL achieves the consistent state-of-theart on all benchmarks. The results in Table 1 are a convincing demonstration of the powerful ability of our approach for time series anomaly detection.

It is worth noting the improvement between AnomalyBERT [10] and our approach. We achieve our self-supervised representation modeling based on synthetic anomalous segments. This idea is come from AnomalyBERT and we also construct our network based on the code provided by it. However, the experiment results show our TDSRL is better that AnomalyBERT, which demonstrates the effectiveness of the new components or methods we applied on it. More experiments will follow to further verify their contributions.

E. Model Analysis

Ablation Studies

Table 2 shows the ablation study of the Frequency-based Degradation for synthetic anomalies generation and the contrastive learning branch (i.e., Pretext task 2) for self-supervised representation modeling. We use SwaT dataset to compare TDSRL with its different variants and conduct the

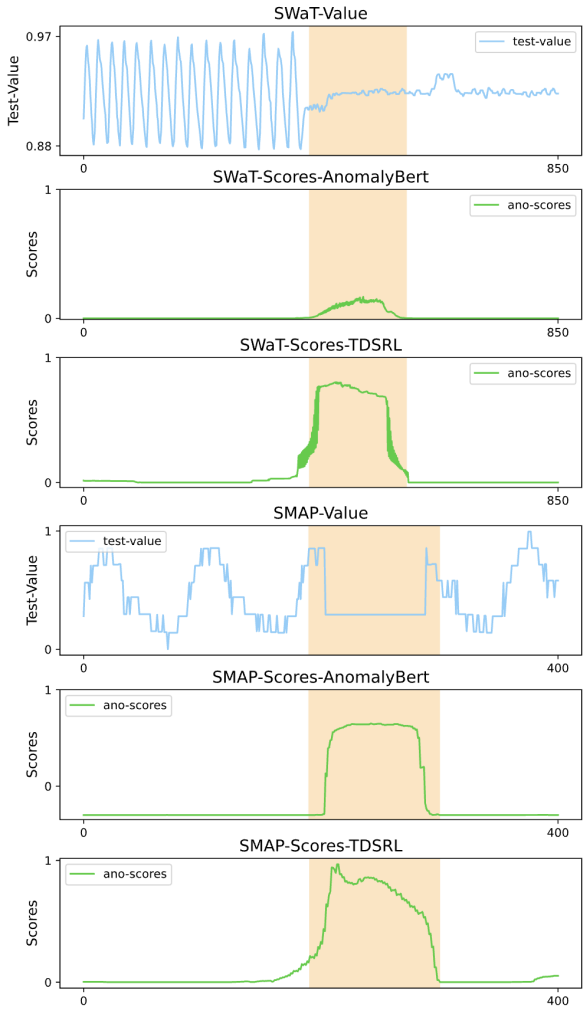

Fig. 3. The anomalous samples in test set and the predicted anomaly scores output from models. The areas marked in orange represent the ground true anomaly segments in test set. Our TDSPL has higher sensitivity to anomaly boundaries and can detect the occurrence of anomalies earlier.

experiments in the same setting. The results show that our TDSRL is the best and further demonstrate the effect of each part. In particular, there is an obvious drop in performance when we eliminate the Pretext task 2. It confirms the great value of this contrastive learning branch for improving anomaly detection accuracy.

Anomaly Detection Visualization

We investigate how TDSRL works by visualizing the anomalous time series samples in test set and the predicted anomaly scores output from the models in Figure 3. We make a comparison with AnomalyBERT [10] using the samples in SWaT and SWAP. It can be seen that the anomaly score produced by TDSRL not only maintains a higher level throughout the anomaly interval, but also has higher values at the beginning and end of the anomaly. On the one hand, these results show the stronger detection ability of TDSRL for various real-world time series anomalies. On the other hand, they verify that our methods can enhance the model's sensitivity to the boundaries between normal and anomalous intervals, thus identifying the occurrence of anomalies more promptly. It's quite significant in many application scenarios. For example, the timely detection of mechanical failures in IoT systems [40] and financial anomalies in transaction markets [41] (or even detecting before they occur) will be very helpful in averting major threats to the safety of people and property [42].

The comparison between anomaly and adjacent normal segments in embedding space

In order to further demonstrate the contribution of our proposed contrastive pretext task, we use the t-SNE algorithm [43] to visualize the high-dimensional features of different segments output by the representation encoder (i.e., the Transformer Body in TDSRL) before and after adding Pretext task 2. Specifically, we randomly select several abnormal segments and their adjacent equal-length normal segments from the validation set to form multiple (S_1, S_2, S_3) triples as shown in Figure 1. Then their representation results from Transformer Body are visualized into two-dimensional space by t-SNE, where each point represents a latent feature and the distance represents their similarity. As shown in Figure 4, our contrastive representation learning can better separate anomalies from adjacent normal intervals in the embedding space. This will help the network to better discriminate between these three locally adjacent segments in the downstream detection module, resulting in more accurate predictions of the location and extent of anomalous segments.

Parameter Sensitivity We also investigate the parameter sensitivity of our TDSRL. Figure 5(a) displays the performance under different proportion threshold Ψ. As shown in Algorithm 1, It is equal to the probability of using Time-based Degradation in the Anomaly Synthesis module. Meanwhile, the proportion of Frequency-based Degradation is equal to $1 - \Psi$. The model performs best when the value of Ψ is near 0.5. This indicates that these two data degradation mechanisms used for anomaly synthesis are of similar importance. Figure 5(b) shows the performance under different anomaly window size l_a, which represents the length of synthetic abnormal segments $X_s^{'}$. The F1-Scores remain stable over a wide range. demonstrating that TDSRL is robust with different anomaly window sizes. Figure 5(c) and Figure 5(d) are used to study

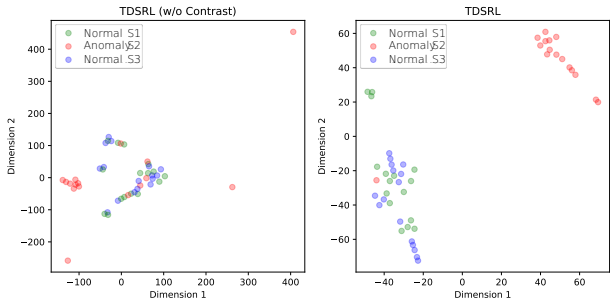

Fig. 4. The visualization of latent features by t-SNE. The three different colours represent the anomaly segments and the adjacent normal segments before and after them. (w/o Contrast) represents the TDSRL variant without the contrastive learning branch.

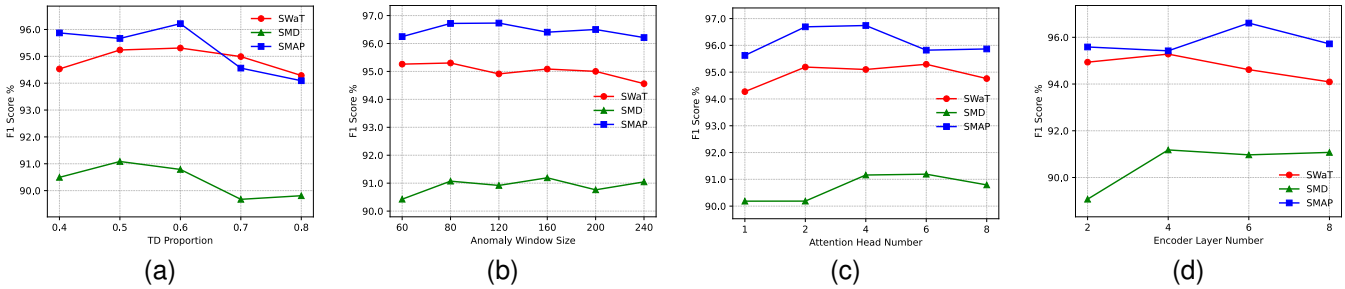

Fig. 5. The study of Parameter Sensitivity for main hyper-parameters in TDSRL. (a) Sensitivity to the proportion between time and frequency based degradation in synthetic anomalies. (b) Sensitivity to the anomaly window size l_a. (c) Sensitivity to the attention head number of Transformer Body. (d) Sensitivity to the encoder layer number of Transformer Body.

the network performance with different numbers of attention heads or encoder layers in the shared Transformer Body of TDSRL, since the performance of many deep neural networks is affected by them. It can be seen that for some datasets, TDSRL tends to achieve the best performance with a smaller number of attention heads and encoder layers than the original setting in the baseline model [10].

V. Conclusion

In this paper, we propose TDSRL to achieve anomaly detection through dual self-supervised representation learning of time series based on synthetic anomaly segments.The experiments on three representative real-world datasets demonstrates the excellent performance of TDSRL and validate the effectiveness of each proposed components. First, we propose a data degradation approach based not only on time but also on frequency to generate synthetic anomaly for self-supervised learning in this task. It allows more natural mimicking of real-world anomalies and enhance the model's generalization ability across different anomalies. Second, we desire a novel contrastive pretext task based on locally contiguous subsegments. It improve model's discrimination ability between anomaly and adjacent normal intervals, resulting in more accurate predictions of the location and extent of anomalous segments. Third, our dual representation learning scheme can not only characterize the relationship between anomalies and the entire time series globally, but also enhance the model's sensitivity to the boundaries between normal and anomalous intervals locally. As a result, our TDSRL can detect the occurrence of anomalies in a more timely manner or even in advance.

References

[1] X. Zhang, Z. Zhao, T. Tsiligkaridis, and M. Zitnik, "Self-supervised contrastive pre-training for time series via time-frequency consistency," *Advances in Neural Information Processing Systems*, vol. 35, pp. 3988–4003, 2022.

[2] Z. Chen, D. Chen, X. Zhang, Z. Yuan, and X. Cheng, "Learning graph structures with transformer for multivariate time-series anomaly detection in iot," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9179–9189, 2021.

[3] H. Zhu, C. Yi, S. Rho, S. Liu, and F. Jiang, "An interpretable multivariate time-series anomaly detection method in cyber-physical systems based on adaptive mask," *IEEE Internet of Things Journal*, 2023.

[4] M. Kacperczyk and E. S. Pagnotta, "Legal risk and insider trading," *The Journal of Finance*, vol. 79, no. 1, pp. 305–355, 2024.

[5] Z. Z. Darban, G. I. Webb, S. Pan, C. C. Aggarwal, and M. Salehi, "Deep learning for time series anomaly detection: A survey," *arXiv preprint arXiv:2211.05244*, 2022.

[6] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," *arXiv preprint arXiv:2110.02642*, 2021.

[7] J. Fan, Z. Liu, H. Wu, J. Wu, Z. Si, P. Hao, and T. H. Luan, "Luad: A lightweight unsupervised anomaly detection scheme for multivariate time series data," *Neurocomputing*, vol. 557, p. 126644, 2023.

[8] D. Huang, L. Shen, Z. Yu, Z. Zheng, M. Huang, and Q. Ma, "Efficient time series anomaly detection by multiresolution self-supervised discriminative network," *Neurocomputing*, vol. 491, pp. 261–272, 2022.

[9] C. Xiao, S. Chen, F. Zhou, and J. Wu, "Self-supervised few-shot time-series segmentation for activity recognition," *IEEE Transactions on Mobile Computing*, 2022.

[10] Y. Jeong, E. Yang, J. H. Ryu, I. Park, and M. Kang, "Anomalybert: Self-supervised transformer for time series anomaly detection using data degradation scheme," *arXiv preprint arXiv:2305.04468*, 2023.

[11] S. Tuli, G. Casale, and N. R. Jennings, "Tranad: Deep transformer networks for anomaly detection in multivariate time series data," *arXiv preprint arXiv:2201.07284*, 2022.

[12] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, and S.-K. Ng, "Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks," in *International conference on artificial neural networks*, pp. 703–716, Springer, 2019.

[13] H. Song, D. Rajan, J. Thiagarajan, and A. Spanias, "Attend and diagnose: Clinical time series analysis using attention models," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[14] Y. Li, X. Peng, J. Zhang, Z. Li, and M. Wen, "Dct-gan: dilated convolutional transformer-based gan for time series anomaly detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 4, pp. 3632–3644, 2021.

[15] M. Sakurada and T. Yairi, "Anomaly detection using autoencoders with nonlinear dimensionality reduction," in *Proceedings of the MLSDA 2014 2nd workshop on machine learning for sensory data analysis*, pp. 4–11, 2014.

[16] E. Eldele, M. Ragab, Z. Chen, M. Wu, C.-K. Kwoh, X. Li, and C. Guan, "Self-supervised contrastive representation learning for semi-supervised time-series classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[17] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3733–3742, 2018.

[18] Y. Yang, C. Zhang, T. Zhou, Q. Wen, and L. Sun, "Dcdetector: Dual attention contrastive representation learning for time series anomaly detection," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3033–3045, 2023.

[19] R. Soklaski, M. Yee, and T. Tsiligkaridis, "Fourier-based augmentations for improved robustness and uncertainty calibration," *arXiv preprint arXiv:2202.12412*, 2022.

[20] H. J. Nussbaumer and H. J. Nussbaumer, *The fast Fourier transform*. Springer, 1982.

[21] Q. Wang, B. Li, T. Xiao, J. Zhu, C. Li, D. F. Wong, and L. S. Chao, "Learning deep transformer models for machine translation," *arXiv preprint arXiv:1906.01787*, 2019.

[22] V. Balntas, E. Riba, D. Ponsa, and K. Mikolajczyk, "Learning local feature descriptors with triplets and shallow convolutional neural networks.," in *Bmvc*, vol. 1, p. 3, 2016.

[23] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 387–395, 2018.

[24] A. P. Mathur and N. O. Tippenhauer, "Swat: A water treatment testbed for research and training on ics security," in *2016 international workshop on cyber-physical systems for smart water networks (CySWater)*, pp. 31–36, IEEE, 2016.

[25] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2828–2837, 2019.

[26] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, "Deep one-class classification," in *International conference on machine learning*, pp. 4393–4402, PMLR, 2018.

[27] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," in *International conference on learning representations*, 2018.

[28] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 387–395, 2018.

[29] D. Park, Y. Hoshi, and C. C. Kemp, "A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1544–1551, 2018.

[30] S. Tariq, S. Lee, Y. Shin, M. S. Lee, O. Jung, D. Chung, and S. S. Woo, "Detecting anomalies in space using multivariate convolutional lstm with mixtures of probabilistic pca," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2123–2133, 2019.

[31] B. Zhou, S. Liu, B. Hooi, X. Cheng, and J. Ye, "Beatgan: Anomalous rhythm detection using adversarially generated time series.," in *IJCAI*, vol. 2019, pp. 4433–4439, 2019.

[32] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2828–2837, 2019.

[33] Y. Shin, S. Lee, S. Tariq, M. S. Lee, O. Jung, D. Chung, and S. S. Woo, "Itad: integrative tensor-based anomaly detection system for reducing false positives of satellite systems," in *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 2733–2740, 2020.

[34] L. Shen, Z. Li, and J. Kwok, "Timeseries anomaly detection using temporal hierarchical one-class network," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13016–13026, 2020.

[35] Z. Li, Y. Zhao, J. Han, Y. Su, R. Jiao, X. Wen, and D. Pei, "Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 3220–3230, 2021.

[36] S. Deldari, D. V. Smith, H. Xue, and F. D. Salim, "Time series change point detection with self-supervised contrastive predictive coding," in *Proceedings of the Web Conference 2021*, pp. 3124–3135, 2021.

[37] Z. Ji, Y. Wang, K. Yan, X. Xie, Y. Xiang, and J. Huang, "A space-embedding strategy for anomaly detection in multivariate time series," *Expert Systems with Applications*, vol. 206, p. 117892, 2022.

[38] M. Kang and B. Lee, "Tictok: Time-series anomaly detection with contrastive tokenization," *IEEE Access*, 2023.

[39] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng, *et al.*, "Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications," in *Proceedings of the 2018 world wide web conference*, pp. 187–196, 2018.

[40] Y. Jin, L. Hou, and Y. Chen, "A time series transformer based method for the rotating machinery fault diagnosis," *Neurocomputing*, vol. 494, pp. 379–395, 2022.

[41] S. Khodabandehlou and S. A. H. Golpayegani, "Market manipulation detection: A systematic literature review," *Expert Systems with Applications*, vol. 210, p. 118330, 2022.

[42] H. Kazemian and S. Shrestha, "Comparisons of machine learning techniques for detecting fraudulent criminal identities," *Expert Systems with Applications*, vol. 229, p. 120591, 2023.

[43] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

APPENDIX

Here we present in detail the feedforward process of representation learning corresponding to the two branches in our TDSRL (i.e., Pretext Task 1 and 2), as shown in Algorithm 2 and 3. At the same time, we also demonstrate the complete logic of dual self-supervised training process for TDSRL, as shown in Algorithm 4.

Algorithm 2 Representation Learning Branch 1: $B^1_{RL}()$

Input: Degraded time series samples $X^{'} \in \mathbb{R}^{T\times d}$

Globle Linear Embedding $z^G = g^G_l(X^{'})$

▷ g^G_l: Globle Patch-Linear module

Global representation result $H^G = g_{rep}(z^G)$

▷ g_{rep}: Transformer-Representation encoder

predicted anomaly scores $Y = g_{det}(H^G)$

▷ g_{det}: Detection Head

return Y

Algorithm 3 Representation Learning Branch 2: $B^2_{RL}()$

Input: Degraded time series samples $X^{'} \in \mathbb{R}^{T\times d}$, the index of synthetic anomaly segment $[t_a, t_a + l_a]$

anomaly segment $S_2 = X^{'}[t_a : t_a + l_a]$

adjacent segment ahead $S_1 = X^{'}[t_a - l_a - 1 : t_a - 1]$

adjacent segment at the back

$S_3 = X^{'}[t_a + l_a + 1 : t_a + 2 \times l_a + 1]$

for all $S_i \in S_1, S_2, S_3$ **do**

  Local Linear Embedding $z^L_i = g^L_l(S_i)$

  ▷ g^L_l: Local Patch-Linear module

  Upsample $z^L_i \in \mathbb{R}^{N^L\times d_E}$ to $z^L_i \in \mathbb{R}^{N\times d_E}$ by interpolation

  Local representation result $H^L_i = g_{rep}(z^L_i)$

  ▷ g_{rep}: Transformer-Representation encoder

end for

return H^L_1,H^L_2,H^L_3

Algorithm 4 Dual Self-supervised Training For TDSRL

Input: Degraded time series samples $X^{'} \in \mathbb{R}^{T\times d}$, Pseudo Labels 1: $\hat{Y} \in \{0,1\}^N$, Pseudo Labels 2: $\langle positive, negative\rangle$ pairs

repeat

  # Pretext Task 1

  $Y = B^1_{RL}(X^{'})$

  $L_B = -\frac{1}{N}\sum_{i=1}^{N}\left[\hat{Y}_i log(Y_i) + (1-\hat{Y}_i) log(1-Y_i)\right]$

  # Pretext Task 2

  $\left\{H^L_1, H^L_2, H^L_3\right\} = B^2_{RL}(X^{'})$

  $D_{pos} = d(H^L_1, H^L_3)$

  $D_{neg1} = d(H^L_1, H^L_2)$

  $D_{neg2} = d(H^L_2, H^L_3)$

  calculate contrastive loss according to Pseudo Labels 2

  $L_C = \sum_{Dneg}(Dpos - Dneg + \delta)$,

  $Dneg \in \{Dneg1, Dneg2\}$

  # synchronous training

  minimize the total loss

  $L_{total} = L_B + L_C$

until L_{total} is less than the threshold or does not decrease for multiple consecutive steps