Information Processing and Management 61 (2024) 103569

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/ipm

Reconstruction-based anomaly detection for multivariate time series using contrastive generative adversarial networks

Jiawei Miao [a], Haicheng Tao [a,*], Haoran Xie [b], Jianshan Sun [c], Jie Cao [c]

[a] College of Information Engineering, Nanjing University of Finance & Economic, 3 Wenyuan Road, Nanjing, 210023, Jiangsu, China
[b] Department of Computing and Decision Sciences, Lingnan University, 8 Castle Peak Road, Tuen Mun, Hong Kong, 999077, Hong Kong, China
[c] Hefei University of Technology, 193 Tunxi Road, Hefei, 230009, Anhui, China

ARTICLE INFO

Keywords:
Anomaly detection
Multivariate time series
Generative adversarial networks
Contrastive learning

ABSTRACT

The majority of existing anomaly detection methods for multivariate time series are based on Transformers and Autoencoders owing to their superior capabilities. However, these methods are susceptible to overfitting when there is insufficient data. To address this issue, we propose a novel unsupervised anomaly detection framework, which seamlessly integrates contrastive learning and Generative Adversarial Networks. More concretely, we utilize data augmentation techniques that incorporate geometric distribution masks to expand our training data, thereby enhancing its diversity. Then, a Transformer-based Autoencoder is trained within a Generative Adversarial Network framework to capture the underlying distribution for normal points. Additionally, we incorporate a contrastive loss into our discriminator to effectively regulate the GAN and ensure good generalization. Finally, anomalies are detected based on reconstruction errors. Numerous experiments on five real-world datasets demonstrated that our proposed method can effectively mitigates overfitting issues and obtains superior performance compared to state-of-the-art approaches. In particular, our model could achieve an average improvement of 9.28% in Precision, 11.33% in Recall, and 11.73% in F1-score.

1. Introduction

Multivariate time series (MTS) is commonly present across a range of domains, inclusive of finance (Cheng, Yang, Xiang, & Liu, 2022; Jiang, Wu, Zhao, Zhu, & Zhang, 2023), electrical power (Wang et al., 2022), and supply chains (Nguyen, Tran, Thomassey, & Hamad, 2021). The analysis of these data can aid in making informed decisions and predictions in information management systems (Duan et al., 2022). Specifically, the detection of anomalies in MTS, which involves identifying and revealing uncommon patterns or events, is crucial for various areas, such as financial fraud prevention (Cheng, Wang, Zhang, & Zhang, 2022) and predictive maintenance (Han, Li, & Shi, 2022). For instance, detecting anomalies in financial time series data can help identify fraudulent activities and mitigate financial risk, while in power grids, it can aid in predicting and preventing power outages. As time series data becomes increasingly complex and diverse, how to precisely identify anomalies remains a challenging task. To address this challenge, a plethora of anomaly detection methodologies have been put forth for MTS data by both academia and industry (Maciąg, Kryszkiewicz, Bembenik, Lobo, & Del Ser, 2021).

Most of earlier studies are primarily focused on statistical methods, such as ARIMA and PCA. Nevertheless, these methods may not yield effective results when applied to MTS data with high volatility and complex patterns. To tackle the aforementioned challenge, particularly when anomaly labels are scarce, various studies have been conducted on clustering-based and One-Class

* Corresponding author.
E-mail address: haicheng.tao@nufe.edu.cn (H. Tao).

https://doi.org/10.1016/j.ipm.2023.103569
Received 6 July 2023; Received in revised form 30 September 2023; Accepted 31 October 2023
Available online 4 November 2023
0306-4573/

classification(OCC) methods. DBSCAN (Chen et al., 2021) and LOF (Breunig, Kriegel, Ng, & Sander, 2000) are two prominent clustering-based methods. These methods are capable of identifying both local and global normal patterns, thereby allowing anomalies or outliers to be characterized as data points that deviate significantly from these normal patterns. As for OCC methods, One-Class SVM (Khreich, Khosravifar, Hamou-Lhadj, & Talhi, 2017) and Isolation Forest (Zhang & Liu, 2022) are widely used. This category of methods construct models based on the distribution of normal MTS data, aiming to accurately classify observations that deviate from this distribution as anomalous instances. However, above methods face challenges when handling with high-dimensional data and may struggle to achieve optimal performance.

Recently, notable advancements have been achieved in the domain of deep learning techniques aimed at detecting anomalies in MTS data. Notably, reconstruction-based deep learning methods have gained widespread attention, especially when labels of anomalies are unavailable (Hundman, Constantinou, Laporte, Colwell, & Söderström, 2018; Ludeña-Choez, Zevallos, & Mayhua-López, 2019; Lv, Wang, & Chen, 2023; Tao et al., 2023). Reconstruction-based methods aim to acquire the underlying representation of normal MTS data and then measure the deviation between the actual MTS data and its reconstructed representation, *i.e.,* reconstruction error. A higher value of the reconstruction error indicates a higher likelihood of the MTS data being classified as anomalous. The choice of representation for MTS data is significant as it bears a direct impact on the overall effectiveness of the detection process (Tao et al., 2023). Hence, a multitude of deep learning methodologies have been introduced with the objective of enhancing the representation of time series data, facilitating a more comprehensive capture of temporal dependencies. Among them, Recurrent Neural Networks (RNNs) (Canizo, Triguero, Conde, & Onieva, 2019), Long Short-Term Memory networks(LSTM) (Abbasimehr, Shabani, & Yousefi, 2020), and Gated Recurrent Units (GRUs) (Thanthawy Sukanda & Adytia, 2022) are commonly employed for representing time series data. More recently, Transformer models (Shao et al., 2023; Xi, Liang, Liu, & Li, 2023; Xu et al., 2018) have emerged as a powerful approach for representing MTS data. Unlike traditional recurrent neural networks, Transformers rely on self-attention mechanisms to establish associations between different data points in the MTS data without the need for recurrent connections. This allows for parallel computation and facilitates modeling long-range dependencies. Kitaev, Kaiser, and Levskaya (2020), Zong et al. (2018). However, these models still suffer from overfitting due to their inherent complexity, thereby impeding their practical applications.

To address above issue, we present a novel method to detect anomalies in MTS based on the Autoencoder (AE) within the Generative Adversarial Network framework. Specifically, to mitigate the problem of overfitting, we incorporated two strategies into our model, *i.e.,* augmenting the dataset with more diverse data and imposing more constraints on the losses. First, we employ a simple trick to augment MTS. Then, we employ an Autoencoder based on the transformer architecture to capture temporal dependencies within the time series. While the Autoencoder is also prone to overfitting, we incorporate Autoencoder into the Generative Adversarial Network framework and impose additional constraints on the reconstruction process. In particular, we also introduce the contrastive constraint into the discriminator to enhance its capacity to capture the underlying distribution of normal patterns. The primary contributions of this manuscript can be succinctly summarized as follows:

- We seamlessly integrate contrastive learning into the GAN framework and propose an unsupervised anomaly detection methods for MTS. This approach effectively mitigates overfitting issues.
- We propose a contrastive constraint by utilizing reconstructed data as positive samples. Additionally, incorporating more negative samples further enhances the representation ability of normal data.
- A comprehensive set of experiments is performed on five distinct real-world datasets and the obtained results significantly corroborate the effectiveness of our proposed approach.

Overall, the proposed method exhibits considerable potential in facilitating the detection of anomalies in MTS, which could have far-reaching implications for numerous practical applications.

The subsequent sections of this manuscript are organized as follows: Section 2 presents an in-depth discussion on the related work. Research objectives are provided in the Section 3. Section 4 outlines our proposed method. The experimental results and corresponding analysis are presented in Section 5. Ultimately, Section 6 encapsulates the concluding remarks of this paper.

2. Related work

Identifying anomalies in MTS data poses a significant challenge due to its multifaceted nature, dynamic behavior, and intricate structures. In response to this challenge, scholars have proposed diverse techniques for the detection of anomalies in MTS data, including statistical approaches and deep learning methods. In this section, we examine the most pertinent literature involved in Transformer-based anomaly detection, GAN-based anomaly detection, as well as contrastive learning methods in MTS data.

2.1. Transformer-based time series anomaly detection

Recently, transformers (Han, Chen, & Liu, 2020) have exhibited formidable capabilities in processing sequential data, including natural language (Devlin, Chang, Lee, & Toutanova, 2019), audio processing (Chen et al., 2022) and computer vision (Huang, 2023), demonstrating their efficacy in these fields.

In time series analysis, transformers leverage the self-attention mechanism to effectively capture reliable long-range temporal dependencies. One classic method of this category is Deep Transformer Models, which uses a multi-headed self-attention mechanism to capture long-term dependencies in time series and uses residual linking to accelerate training and improve model performance.

TranAD (Tuli, Casale, & Jennings, 2022) is a deep-learning based model, specifically designed for identifying anomalous patterns present in multivariate time series. The model harnesses the transformative properties of self-attention encoders, facilitating the comprehensive capture of intricate temporal dependencies. Additionally, this approach harnesses focus score-based self-regularization techniques that enhance robustness in multi-modal feature extraction. To further ensure the stability and efficacy of the model, adversarial training has been employed extensively. Furthermore, given the challenge of limited data, it has utilized model-agnostic meta-learning (MAML) as an effective technique for efficient model training.

Anomaly Transformer (Xu, Wu, Wang, & Long, 2022) algorithm is a method for detecting anomalies in time series that is built upon the robust Transformer model. At the core of this approach lies the concept of exploiting the self-attention mechanism intrinsic to the Transformer model, and leveraging it to scrutinize the distribution of self-attention weights at every time step of the time series. Through this, the method captures rich associations between sub-sequences and the entire time sequence, thereby facilitating the identification of anomalous data-points.

The UTAD in ECG Signals (Alamr & Artoli, 2023) is a novel method for detecting anomalies in electrocardiogram (ECG) signals, utilizing a transformer-based architecture that has demonstrated impressive capabilities on two benchmark datasets, namely ECG5000 and MIT-BIH Arrhythmia. The proposed model comprises two key components: an embedding layer and a conventional transformer encoder.

Notwithstanding their efficacy, some challenges still hinder the practical application of these methods. Notably, there is a risk of overfitting to anomalous patterns during reconstruction, as well as non-robustness while training on contaminated data.

2.2. GAN-based time series anomaly detection

Generative Adversarial Networks (GANs) have gained significant attention for unsupervised anomaly detection in time series (Han et al., 2020). Typically, GANs consist of two key components, *i.e.,* Generators and Discriminators, which are trained in a game-theoretic framework. In order to detect anomalies, the Generator is trained as an Autoencoder. And reconstructed data that closely adheres to the underlying structure of the original data can be generated by the Generator. Based on the difference between the reconstructed data and the original data, anomalies can be effectively detected.

The field of GAN-based anomaly detection has witnessed rapid advancements, with methods such as GANomaly (Akcay, Abarghouei, & Breckon, 2018), MAD-GAN (Li et al., 2019), TadGAN (Geiger, Liu, Alnegheimish, Cuesta-Infante, & Veeramachaneni, 2020), and BeatGAN (Zhou, Liu, Hooi, Cheng, & Ye, 2019) emerging as prominent contributors. These methodologies harness the capabilities of GAN architectures to generate synthetic samples that obey the distribution of original data. The GAN framework inherently encompasses global regularization mechanisms during the training process, effectively addressing the issue of overfitting commonly encountered in anomaly detection tasks.

However, a major challenge arises when the training set includes anomalous instances. In such scenarios, the model tends to learn the contaminated mixture distribution, leading to a compromise in its effectiveness for detecting anomalies. This phenomenon significantly impedes the model's capacity to precisely discriminate between normal and abnormal patterns, thereby impacting its performance in real-world scenarios. FGANomaly (Du et al., 2021) addresses this problem to a large extent by using pseudo-labeling and filters. But the effectiveness of FGANomaly is to some extent limited by the accuracy of the pseudo-labels.

2.3. Contrastive learning in time series

Contrastive learning (Chopra, Hadsell, & LeCun, 2005; Sohn, 2016) is a self-supervised learning methodology that seeks to learn a data representation that maximizes the similarity between positive pairs while minimizing the similarity between negative pairs. Its potential has been recognized in various domains, including time series forecasting and anomaly detection (Hao, Wang, Alexander, Yuan, & Zhang, 2023; Pöppelbaum, Chadha, & Schwung, 2022; Woo, Liu, Sahoo, Kumar, & Hoi, 2022; Yang, Zhang, & Cui, 2022).

Lee, Byun, and Baek (2023) proposed a contrastive learning-based anomaly detection method, which introduces a newly defined objective function for simultaneous learning of the OCC model and contrastive learning. Furthermore, by applying appropriate data augmentation techniques for periodic time series data in contrastive learning, it is possible to extract features that preserve temporal characteristics.

This section presents a comprehensive overview of the most significant research endeavors in the fields of GAN-based time series anomaly detection, Transformer-based anomaly detection, and contrastive learning in time series. These cutting-edge techniques have exhibited remarkable capabilities in detecting anomalies within time series data.

3. Research objectives

Transformers and Autoencoders can yield impressive results in MTS anomaly detection. However, overfitting remains a significant concern. Hence, the main research problem is how to mitigate the risk of overfitting when using a complex model structure like a transformer in combination with a GAN. In this paper, we adopt two strategies to address overfitting, *i.e.*, increasing training data and applying regularization techniques. Consequently, the primary research objectives are to propose simple yet effective data augmentation methods for expanding the training data and develop regularization methods for improving model generalization. Additionally, another research objective is to propose a methodology for seamlessly combining these strategies to achieve optimal results. In following sections, we will delve into the solutions for these research objectives.

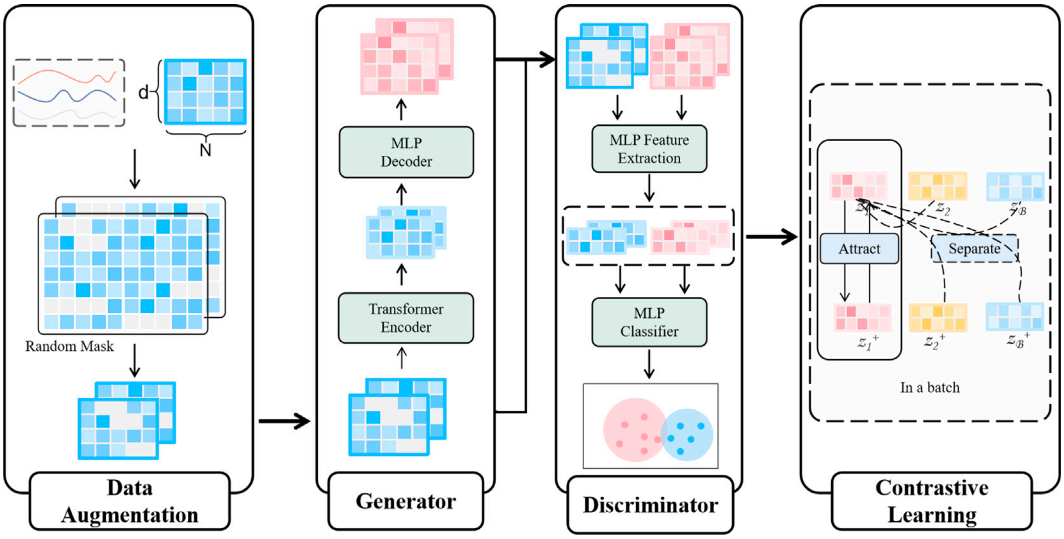

Fig. 1. Overview of the Proposed Model. The input MTS data are augmented via random masks based on a geometric distribution. Then, the augmented data are passed through a Generator to produce the reconstructed data. The Generator consists of a transformer encoder and an MLP Decoder. For discriminator, the augmented data ("Real") and the reconstructed data ("Fake") are passed through an MLP to obtain their latent representations and an simple MLP classifier to identify "Real" or "Fake". Moreover, a contrastive constraint is imposed for latent representations. Positive samples are from the same MTS representations in a batch. The model is trained within the GAN framework.

4. Method

4.1. Problem definition

Definition 1 (*Multivariate Time Series*)**.** A multivariate time series $\mathcal{X} \in \mathbb{R}^{N\times d}$ is composed of N consecutive observation points $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^d$. These observation points generally have equally spaced time intervals, which can be measured in seconds, minutes, or hours.

Definition 2 (*Sliding Window*)**.** A sliding window is a segment sampled from the MTS $\mathcal{X}$. It can be represented as $\mathcal{X}_{i:i+w-1} = \{\mathbf{x}_i, \mathbf{x}_{i+1}, \ldots, \mathbf{x}_{i+w-1}\}$, where the length of the sliding window is w.

Problem statement. Given a multivariate time series $\mathcal{X} \in \mathbb{R}^{N\times d}$, we aim to identify anomalous points without utilizing any known ground truth labels. To detect anomalies, reconstruction errors are calculated using the input segments through sliding windows and their corresponding reconstructed segments. Any points with errors exceeding the specified threshold will be identified as anomalies. In this paper, we optimize the threshold to maximize the F1 score.

4.2. Overview of proposed model

The overview of the proposed model is illustrated in Fig. 1, comprising four main modules, *i.e.,* data augmentation, generator, discriminator, and contrastive learning module.

- **Data Augmentation** This module leverages the properties of MTS and expand the unexplored input space using a novel method called random mask, which follows a geometric distribution.
- **Generator** This module learns the underlying distribution of normal patterns in MTS and reconstructs it precisely using a Transformer-based autoencoder.
- **Discriminator** This module imposes constraints on the reconstructions within the GAN framework to better capture normal patterns in MTS.
- **Contrastive Learning** This module imposes contrastive constraints on representations of MTS to enhance the generalization capability of the model, and facilitates joint training of the Discriminator.

4.3. Data augmentation

Data augmentation (Xiao, Zhang, Chen, & Yin, 2021) is a critical technique for improving both the quantity and quality of training data, thereby reducing model overfitting. Conventional time series data augmentation approaches incorporate techniques such as window cropping, window warping, and flipping in the time domain, along with some methods for augmenting data in the frequency domain. However, these

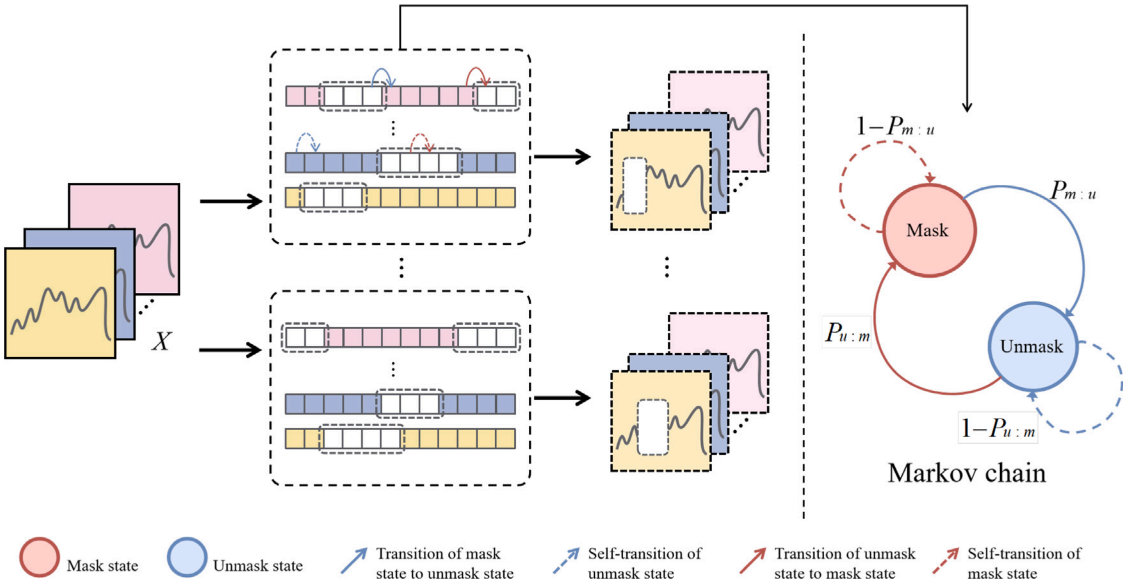

Fig. 2. Data augmentation model.

transformations may have the potential to introduce artificial patterns or noise into the data, which increases the risk of overfitting . In this sense, we adopt a simple but effective method known as random mask that adheres to a geometric distribution.

As shown in Fig. 2, the binary noise mask matrix construction process, denoted as M, utilizes a Markov chain consisting of two states: "Mask" and "Unmasked". The state transition matrix in the Markov chain can be expressed as follows:

$$P = \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} = \begin{bmatrix} p_{m:m} & p_{m:u} \\ p_{u:m} & p_{u:u} \end{bmatrix} \tag{1}$$

Given the masked proportion r and the mean length of the mask segment l_m, we can obtain the transition probability of the "Mask" state $p_{m:u} = \frac{1}{l_m}$ based on the geometric distribution. And the transition probability of the "Unmask" state is then determined as $p_{u:m} = p_{m:u} * \frac{r}{1-r}$. The complete transition probability matrix can be formulated as follows:

$$\begin{aligned} P(s_{t+1} = 0 | s_t = 0) &= p_{m:m} = 1 - \frac{1}{l_m} \\ P(s_{t+1} = 1 | s_t = 0) &= p_{m:u} = \frac{1}{l_m} \\ P(s_{t+1} = 0 | s_t = 1) &= p_{u:m} = \frac{1}{l_m} * \frac{r}{1-r} \\ P(s_{t+1} = 1 | s_t = 1) &= p_{u:u} = 1 - \frac{1}{l_m} * \frac{r}{1-r} \end{aligned} \tag{2}$$

Considering that the MTS $\mathcal{X} \in \mathbb{R}^{N \times d}$, we can obtain positive pairs by utilizing two independent mask operations randomly applied to each dimension, which can be formulated as follows:

$$\hat{\mathcal{X}} = M \odot \mathcal{X} \tag{3}$$

where $M \in \{0,1\}^{N \times d}$ is a binary noise mask matrix following a geometric distribution, $\hat{\mathcal{X}}$ is a transformed MTS after data augmentation. In our experiments, we generate two separate masks for each training sample to obtain a positive sample pair.

4.4. Generator

As illustrated in Fig. 1, our model is within the GAN framework, in which Generator G endeavors to reconstruct the original MTS including augmented MTS, and Discriminator D aims to distinguish original MTS and reconstructed MTS. In particular, the generator G follows an encoder–decoder architecture. The encoder f_{enc} maps the augmented MTS into latent representation $\mathcal{Z}$, while the decoder f_{dec} reconstructs the MTS from the $\mathcal{Z}$, which can be formulated as follows:

$$f_{enc} : \hat{\mathcal{X}} \to \mathcal{Z}, \qquad f_{dec} : \mathcal{Z} \to \hat{\mathcal{X}}'. \tag{4}$$

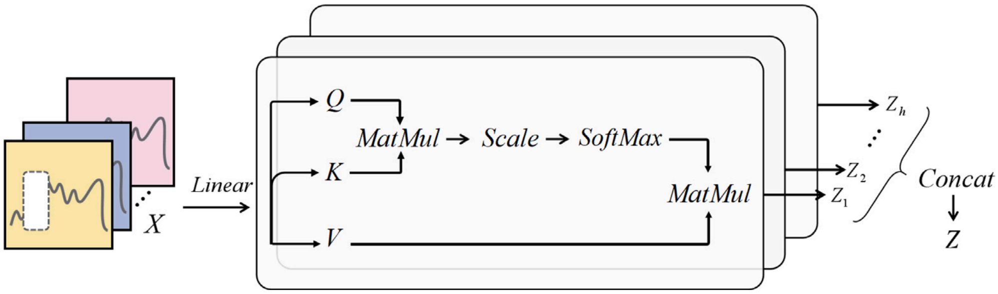

Fig. 3. Structure of the transformer encoder.

In order to capture long-term dependencies and patterns in the normal MTS, the encoder f_{enc} is based on the Transformer as shown in Fig. 3, which can be formulated as follows:

$$\begin{aligned} Q_i,\ K_i,\ V_i &= \hat{\mathcal{X}}W_i^Q,\ \hat{\mathcal{X}}W_i^K,\ \hat{\mathcal{X}}W_i^V \\ \mathcal{Z}_i &= \operatorname{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right)V_i \\ \mathcal{Z} &= \operatorname{Concat}(\mathcal{Z}_1,\dots,\mathcal{Z}_h)W^O \end{aligned} \tag{5}$$

Where W_i^Q, W_i^K, W_i^V represent the parameter matrices for query, key, value of Multi-Head attention, h is the number of the head.

Our decode is a two-layer MLP which is formulated as follows:

$$\begin{aligned} \mathcal{H}_g &= \sigma(W_{g1}\cdot\mathcal{Z}+\mathbf{b}_{g1}) \\ \hat{\mathcal{X}}' &= \sigma(W_{g2}\cdot\mathcal{H}_g+\mathbf{b}_{g2}) \end{aligned} \tag{6}$$

Where W_{g1} and W_{g2} are the weight matrices for the first and second layer respectively, $\mathbf{b}_{g1}$ and $\mathbf{b}_{g2}$ are the bias vectors, and σ is the Sigmoid activation function.

Then, the reconstruction loss can be formulated as:

$$L_{\text{rec}} = \frac{1}{N}\left\|\hat{\mathcal{X}} - \hat{\mathcal{X}}'\right\|_2 \tag{7}$$

4.5. Discriminator

In this section, We employ the Discriminator D to impose constraints on the Generator G to regularize the reconstruction errors. In particular, we denote the original and reconstructed MTS data as real and fake, respectively. And the Discriminator D is a binary classifier, assigning 0 or 1 to the input MTS, which can be formulated as follows:

$$f_{dis} : \mathcal{X}_{dis} \to [0, 1] \tag{8}$$

Where $\mathcal{X}_{dis}$ can be original augmented MTS data and reconstructed augmented MTS data.

In our implementation, f_{dis} is constructed as a three-layer MLP. The first two layers are dedicated to feature extraction, while the final layer serves the purpose of classification, which can be formulated as follows:

$$\begin{aligned} \mathcal{H}_{d1} &= \sigma(W_{d1}\cdot\mathcal{X}_{dis}+\mathbf{b}_{d1}) \\ \mathcal{H}_{d2} &= \sigma(W_{d2}\cdot\mathcal{H}_{d1}+\mathbf{b}_{d2}) \\ P(Y|\mathcal{X}_{dis}) &= \operatorname{softmax}(W_{d3}\cdot\mathcal{H}_{d2}+\mathbf{b}_{d3}) \end{aligned} \tag{9}$$

Where W_{d1}, W_{d2} and W_{d3} are the weight matrices for the first, second and last layer respectively, $\mathbf{b}_{d1}$, $\mathbf{b}_{d2}$ and $\mathbf{b}_{d3}$ are the bias vectors, $\mathcal{H}_{d2}$ and $\mathcal{H}_{d2}$ are hidden latent representations for MTS data, Y is the label for real or fake and σ is the Sigmoid activation function.

Within the GAN framework, the Discriminator loss L_{dis} can be formulated as follows:

$$L_{dis} = -\frac{1}{N}\left[\log D\left(\mathcal{X}_{dis}\right) + \log\left(1 - D\left(\mathrm{G}\left(\mathcal{X}_{dis}\right)\right)\right)\right] \tag{10}$$

4.6. Contrastive learning

Moreover, we apply a contrastive constraint on the latent representations of the reconstructed MTS data to further regularize the reconstruction errors. This constraint could enhance the quality of the reconstructed MTS data by encouraging meaningful and distinguishable latent representations.

In a batch, we only consider representations of two reconstructed MTS data as positive pairs and others as negative pairs. As illustrated in Fig. 1, a contrastive constraint can be applied to encourage the latent representations of positive pairs to be closer together (“attract”) while pushing the latent representations of negative pairs further apart (“separate”). The contrastive constraint can be defined as:

$$L_{con} = -\log\frac{\exp(cos(\mathcal{H}_{d2}\cdot\mathcal{H}_{d2}^{+})/\tau)}{\sum_{i=0}^{B}\exp\left(cos(\mathcal{H}_{d2}\cdot\mathcal{H}_{d2}^{i})/\tau\right)} \tag{11}$$

Where τ is the temperature factor and + denotes the positive pairs. B represents the total number of samples in the current training batch.

4.7. Model training

Given the input MTS data $\mathcal{X} \in \mathbb{R}^{N \times d}$, the Generator and the Discriminator are trained alternately. During the training process, we incorporate both the Generator adversarial loss L_{adv} and reconstruction loss L_{rec} to train the Generator, which can be formulated as follows:

$$L_{adv} = -\frac{1}{N} \sum_{i=1}^{N} \log D\left[G\left(\mathcal{X}\right)\right] \tag{12}$$

$$L_G = L_{rec} + \lambda \cdot L_{adv} \tag{13}$$

λ is used to parameterize the trade-off between L_{adv} and L_{rec}.

Moreover, the Discriminator is trained by incorporating both the Discriminator loss L_{dis} and contrastive loss L_{con}, which can be formulated as follows:

$$L_D = L_{dis} + \gamma \cdot L_{con} \tag{14}$$

γ is employed to balance the importance of L_{dis} and L_{con}.

Algorithm 1 shows the whole training process.

Algorithm 1 Model Training

Require: An input MTS: $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$; training epochs I; batch size B; the balance factor of adversarial loss λ, γ;
Ensure: Trained G and D

```
Initialize the parameters for the Generator G and the Discriminator D
while epoch < I do
    Sample B MTS data;
    Generate augmented MTS data by Eq. (3)
    Generate reconstructed MTS data by Eq. (5) and Eq. (6)
    Compute L_dis by Eq. (10)
    Compute L_con by Eq. (11)
    L_D ← L_con + γ · L_dis
    Update parameters of the Discriminator D by L_D
    Compute L_rec by Eq. (7)
    Compute L_adv by Eq. (12)
    L_G ← L_rec + λ · L_adv
    Update parameters of the Generator G by L_G
    epoch = epoch + 1
end while
```

5. Experiments

In this section, extensive experiments were carried out on five real-world datasets in order to evaluate the effectiveness of our proposed approach. Additionally, we performed various ablation studies to investigate the effects of different components in our model. Furthermore, a case study was also introduced to provide additional insights into the performance of the model.

5.1. Experimental setup

5.1.1. Public datasets

The experiments were conducted on five publicly available datasets, which are described as follows:

- **SMAP dataset (Soil Moisture Active Passive)** is a dataset launched by NASA (Hundman et al., 2018) for soil moisture remote sensing research. The dataset collects soil moisture data from around the world for studying the spatio-temporal variations of soil moisture.
- **MSL dataset (Mars Science Laboratory)** is a dataset provided by NASA, is analogous to SMAP in that it comprises sensor and actuator data originating from the Mars rover. The dataset's 55 dimensions encompass telemetry anomaly data extracted from the spacecraft's anomaly detection and resolution system (ISA) reports.
- **PSM dataset (Pooled Server Metrics)** is an internal aggregation of multiple application server nodes at eBay.
- **SWaT dataset (Secure Water Treatment)** is a dataset developed by researchers at Nanyang Technological University in Singapore for water treatment process security research (Goh, Adepu, Junejo, & Mathur, 2016). The dataset simulates a real water treatment plant and generates a large amount of sensor data for studying related network security issues.
- **MIT-BIH Supraventricular Arrhythmia Database (MBA)** consists of electrocardiogram recordings from different patients, encompassing instances of two distinct types of anomalies, namely supraventricular contractions and premature heartbeats. As a widely adopted and sizable dataset in the realm of data administration, it has been a popular option for research purposes.

The detailed statistics of the five datasets are presented in Table 1.

Table 1
Dataset statistics.

Dataset	#Train	#Test	#Dimensions	Anomalies (%)
MSL	58,317	73,729	55	10.72
SMAP	135,183	427,617	25	13.13
PSM	105,984	87,841	25	27.8
SWaT	496,800	449,919	51	11.98
MBA	100,000	100,000	2	0.14

5.1.2. Baselines

We consider 6 baseline anomaly detection methods, which are outlined as follows:

- **Principal component analysis (PCA)** is widely used in dimensionality reduction and feature extraction for time series data. PCA can effectively capture underlying patterns and trends latent within the time series data.
- **MERLIN** (Nakamura, Imamura, Mercer, & Keogh, 2020) identifies anomalies through an iterative process that compares subsequences of varying lengths to their adjacent sequences.
- **MAD-GAN** (Li et al., 2019) employs an LSTM-integrated Generative Adversarial Network to emulate the distribution of time series data. The model combines prediction error with discriminator loss to derive an anomaly score.
- **GDN** (Deng & Hooi, 2021) learns graph relationships among feature patterns, outputs anomaly scores using attention-based prediction and deviation scores.
- **Unsupervised Anomaly Detection (USAD)** (Audibert, Michiardi, Guyard, Marti, & Zuluaga, 2020) is an unsupervised anomaly detection method that leverages adversarial training of autoencoders. The model involves training two autoencoders simultaneously in an adversarial manner and utilizes the reconstruction errors to identify anomalies.
- **TranAD** (Tuli et al., 2022) is a novel approach for detecting and diagnosing anomalies in multivariate time series data. TranAD is built on the transformer that leverages encoders with multi-head attention to enable efficient inference while capturing comprehensive temporal trends inherent in MTS data.
- **BeatGAN** (Zhou et al., 2019) presents a novel multivariate time series anomaly detection methodology grounded GANs. The core architecture of BeatGAN integrates an autoencoder as the primary component, complemented by a discriminator serving as adversarial regularization. This sophisticated amalgamation of techniques showcases the potential of BeatGAN as an effective solution for identifying anomalies in MTS data.
- **FGANomaly** (Du et al., 2021) is a novel GAN-based method for detecting anomalies in multivariate time series data. FGANomaly incorporates a screening process, utilizing pseudo-labeling, to identify prospective anomalous samples prior to training the discriminator component.
- **Anomaly Transformer (AT)** (Xu et al., 2022) introduces an innovative approach utilizing the Transformer model for the detection of anomalies within time series data. The model leverages prior-association and series-association to distinguish normal and abnormal time points.

5.1.3. Evaluation metrics

To assess the efficacy of anomaly detection, we employ the Precision (P), Recall (R), and F1 Score ($F1$) metrics. The definitions of these metrics are as follows: Precision (P) characterizes the ratio of true positive samples in relation to all samples predicted as positive, and it is calculated as

$$P = TP/(TP + FP),$$

where TP represents the count of true positives and FP represents the count of false positives.

R (Recall) measures the proportion of true positive samples among all positive samples, and it is calculated as

$$R = TP/(TP + FN),$$

where FN represents the count of false negatives.

F1 Score is the harmonic mean of P and R, which offers a balanced evaluation by considering both aspects and provides a consolidated measure of performance for anomaly detection, and it is calculated as

$$F1 = 2 * P * R/(P + R).$$

5.1.4. Experimental settings

In our experiments, the sliding window width is set as 200 for MSL, 100 for SMAP, 200 for PSM, 200 for SWaT, and 200 for MBA, respectively. We trained the models using the Adam optimizer with a learning rate of 1e−4. The hyperparameter λ is set as 0.5, and γ is set as 1 (Eqs. (9) and (10)). The Transformer encoder consists of 2 identical layers. Each of these layers is composed of a multi-head self-attention mechanism, with 4 attention heads, and a feed-forward neural network that contains 2 layers.

The training of the methods involved was conducted on a Windows server equipped with a 3.20 GHz AMD Ryzen 7 5800H CPU and an 4 GB Nvidia GeForce RTX 3050 Ti GPU.

5.2. Overview performance

Table 2 shows the precision, recall, and F1 scores for both our model and the baselines on all datasets. Compared to the **9** baseline methods presented in the table, our model can achieve an average increase of 11.73% in F1-score across all five datasets, with an average improvement of 9.28% in Precision and 11.33% in Recall. Further details for each dataset are provided as follows:

Table 2
Overview performance of all methods (%).

Dataset	MSL			SMAP			PSM			SWaT			MBA		
Metrics	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
PCA	**93.66**	84.01	88.57	90.69	57.60	70.45	78.48	99.24	87.65	38.67	32.89	35.54	63.12	89.47	74.02
MERLIN	52.21	48.45	50.26	25.72	99.43	40.87	54.91	98.92	70.62	65.60	25.47	36.69	98.46	49.13	65.55
MAD-GAN	85.16	99.30	91.69	81.57	92.16	86.54	88.63	96.51	92.40	95.93	69.57	80.65	93.96	100	96.89
GDN	92.08	98.92	95.02	74.80	98.91	85.18	90.14	93.69	91.88	96.97	69.57	81.01	88.32	98.92	93.32
USAD	89.40	87.62	88.50	83.93	88.05	85.94	90.60	99.74	94.95	97.52	76.78	85.92	89.53	99.89	94.43
TranAD	90.38	**100**	94.94	80.43	**100**	89.15	92.62	**99.74**	96.05	**99.77**	68.79	81.43	95.69	**100**	97.80
BeatGAN	89.12	86.36	87.71	78.95	93.69	85.56	89.83	93.50	91.60	73.60	92.94	82.14	94.07	96.52	95.28
FGANomaly	90.05	93.60	91.79	76.09	99.93	86.40	94.92	98.78	96.81	98.51	79.54	88.01	96.32	96.14	96.23
AT	91.28	85.89	88.50	**92.59**	97.50	**94.98**	96.82	97.96	97.39	91.48	95.24	93.32	94.31	98.78	96.49
Ours	92.56	99.26	**95.43**	88.18	99.39	93.45	**97.32**	97.79	**97.56**	92.16	**96.52**	**94.29**	**99.34**	99.42	**99.38**

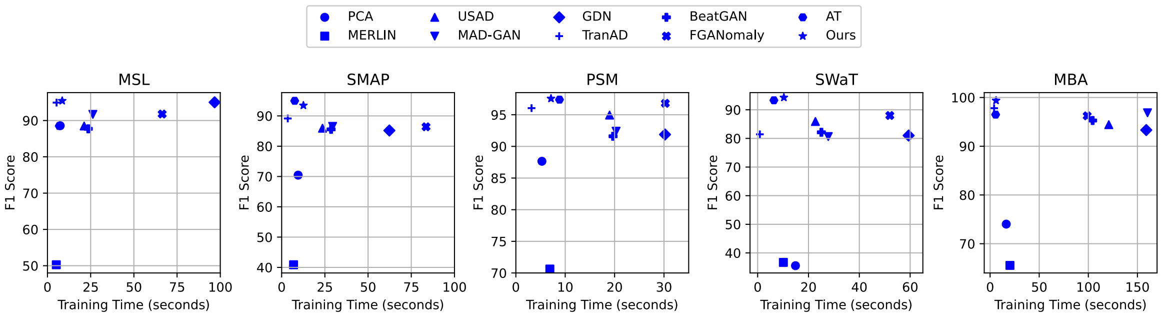

Fig. 4. Performance comparison of computational efficiency and F1 scores.

- On MSL, our model demonstrated an improvement of 6.5% in Precision, 12.99% in Recall, and 9.46% in F1-score over the baselines.
- On SMAP, we attained gains of 12.09% in Precision, 7.47% in Recall, and 12.89% in F1-score compared to the baselines.
- On PSM, our Precision exceeded baselines by 11.22%, with improvements of 0.23% in Recall and 6.52% in F1-score.
- On SWaT, our model surpassed all the baselines, achieving a 7.93% increase in Precision, a 28.65% increase in Recall, and a 20.53% increase in F1-score.
- On MBA, our model achieved improvements of 8.92%, 7.33% and 9.38% in Precision, Recall and F1-score, respectively.

In particular, for the MSL dataset, its low variance poses a challenge for methods to effectively learn normal patterns and exhibit notable distinctions among them. However, our model still outperforms other models with an F1 score of 0.9543. Moreover, our model achieves a significantly higher Recall and Precision scores compared to baselines. This indicates that our model can accurately identify anomalies with fewer false positives and achieve a better balance between Precision and Recall. Similarly, it is indeed challenging to capture the interdependencies between different dimensions for multivariate time series data for the MSL and SWaT datasets, due to their high dimensionality. While most baseline methods tend to exhibit relatively poor performance, our approach achieves an impressive Recall and F1 scores. These results further validate the effectiveness of our approach.

Furthermore, we can observe that methods based on the transformer, *i.e.*, TranAD, FGANomaly, AT and our model, demonstrate notably higher performance in terms of F1 score than other baselines. This observation suggests that capturing long-term dependencies is crucial for effectively distinguishing anomalies from normal patterns. In contrast, methods like PCA, which primarily focus on capturing the overall variance, may not fully capture these intricate temporal relationships.

Considering the efficiency, we present a direct comparison of the training time in each epoch in Table 3, contrasting the baselines with our model. Despite not being the fastest, our model often falls within the top tier in terms of training time, making it acceptable. Furthermore, we provide an in-depth comparison of model performance relative to computational efficiency in Fig. 4. In the case of our model, the findings are compelling. Especially, the effectiveness of our model often exceeds that of faster models in most cases. These results further affirm the consistent and superior performance of our model.

5.3. Ablation study

In this section, we conduct a thorough investigation into the effects of each component within our model. First, We compare our data augmentation component, *i.e.* Geometric mask, with three other widely-used data augmentation techniques. Window warping is a distinctive data augmentation technique applied to time series data. The core idea of this method is to change the time axis in time series data, introducing temporal variations by stretching or compressing different parts of the signal. STFT(Short-Time Fourier Transform) is a commonly employed time–frequency analysis method in signal processing. It divides the signal into small time windows and performs a Fourier transform on each time window to obtain a frequency domain representation of the time series. Bernoulli Mask augmentation is a data augmentation technique similar to the geometrically distributed mask we are employing. Its main change is that an independent Bernoulli distribution is used to produce a more discrete mask matrix, with no change in mask radio.

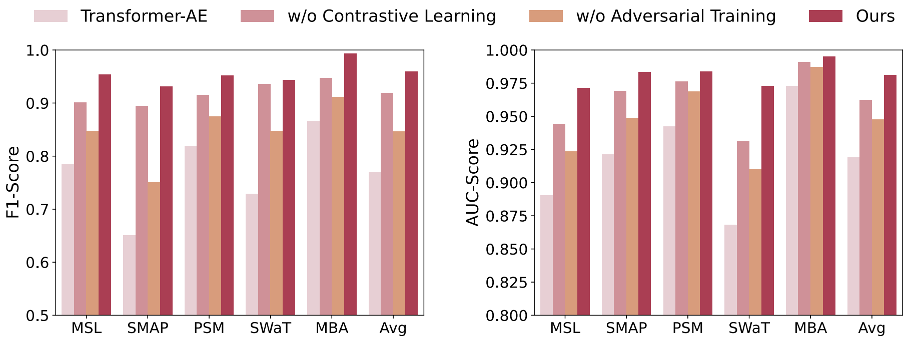

Fig. 5. Ablation studies.

Table 3
Comparison of training times in seconds per epoch.

Method	MSL	SMAP	PSM	SWaT	MBA
PCA	7.38	9.54	5.28	14.88	16.42
MERLIN	5.12	6.89	6.89	10.12	20.19
USAD	21.22	23.63	18.98	22.72	120.86
MAD-GAN	26.27	29.49	20.3	27.79	160.29
GDN	96.71	62.33	30.21	59.4	159.01
TranAD	5.27	3.55	3.18	0.87	4.08
BeatGAN	23.63	28.6	19.69	25.07	104.39
FGANomaly	66.37	83.54	30.27	52.01	98.85
AT	6.83	7.54	8.84	6.39	5.63
Ours	8.53	12.59	7.48	10.32	6.04

Table 4
Performance of different data augmentation methods (%).

Method	MSL		SMAP		PSM		SWaT		MBA	
	F1	AUC	F1	AUC	F1	AUC	F1	AUC	F1	AUC
Window warping	90.45	93.64	90.32	95.63	91.79	97.42	92.28	94.26	95.04	99.01
STFT augmented	95.52	94.96	91.72	97.47	92.39	96.99	92.44	95.74	93.12	98.83
Bernoulli mask	93.84	95.89	91.15	97.19	93.34	98.02	93.88	95.01	95.73	99.43
Geometric mask	95.43	**97.14**	**93.18**	**98.35**	**97.56**	**98.38**	**94.39**	**97.29**	**99.38**	**99.53**

As shown in Table 4, MTS anomaly detection based on geometric mask achieves the highest F1 score and AUC on most of datasets. This indicates that geometric mask could introduce more beneficial perturbations to the MTS data, enabling the model to generalize better and achieve superior performance on various datasets. As a result, the geometric mask technique is a compelling choice for MTS anomaly detection. Second, considering the choice of reconstruction loss can significantly influence how well an anomaly detection model fits the MTS data, especially when combined with specific data augmentation techniques. In our additional experiments, we contrast our use of the MSE loss with the Mean Absolute Error (MAE) loss. The comparison results are as shown in Table 5. Notably, when using the MSE loss with the geometric masking, we achieve results that outperform all other methods in terms of F1 scores. This suggests that the combination of the MSE loss and geometric masking creates an ideal setting for our proposed method. Finally, we delve deeper into key models of our proposed method. The first model, named Transformer-AE, is trained solely with a plain transformer-based autoencoder using mean squared error (MSE) loss. The second model, referred to as w/o Contrastive Learning, is a transformer-based generative adversarial network (GAN) trained without incorporating contrastive learning. The third model, named w/o Adversarial Training, is a transformer-based autoencoder trained exclusively without the incorporation of adversarial training.

We conduct experiments on all five datasets, and the results are shown in Fig. 5. It is evident that all three models are crucial, as removing any one of them results in a decrease in both F1-score and AUC. Particularly, our model with adversarial training and contrastive constraint could bring a remarkable 18.98% (77.01% → 95.99%) average F1-score promotion and a 6.24% promotion in average AUC-score (91.90% → 98.14%). This implies that the two proposed strategies successfully mitigate the issue of overfitting and enhance the model's capacity to generalize effectively.

Table 5
Performance of different losses and data augmentation methods (%).

Method	Recon loss	Data augmentation	MSL	SMAP	PSM	SWaT	MBA	Avg F1
	MAE	Bernoulli mask	82.52	88.76	89.93	76.19	85.14	84.51
	MAE	Geometric mask	89.83	84.24	90.21	80.87	87.56	86.54
	MSE	Bernoulli mask	93.84	91.15	93.34	93.88	95.73	93.58
Final	MSE	Geometric mask	**95.43**	**93.18**	**97.56**	**94.39**	**99.38**	**95.98**

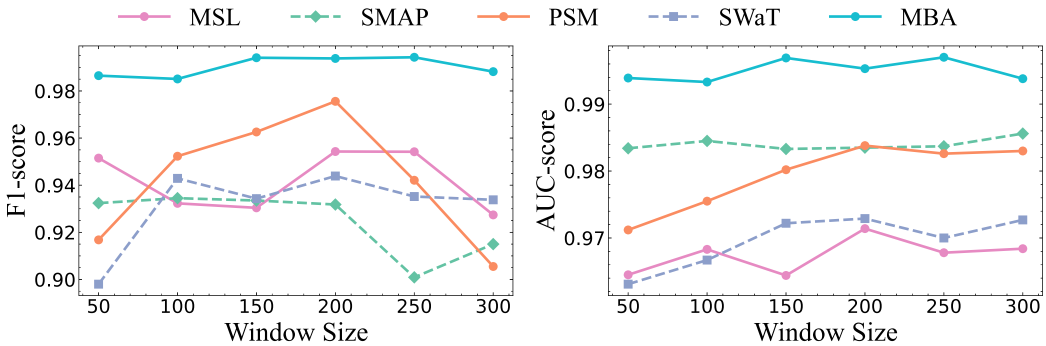

Fig. 6. F1 an AUC scores with different sliding window lengths.

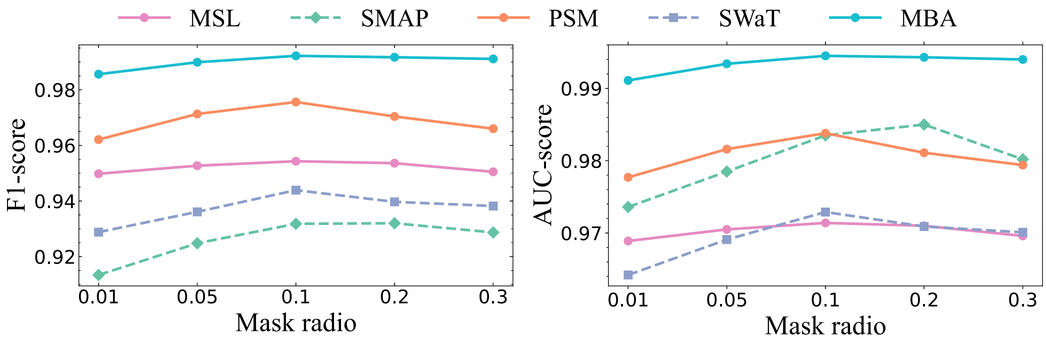

Fig. 7. F1 and AUC scores with different mask radios in data augmentation.

5.4. Parameter sensitivity

In this section, we investigate how the window size (w) and mask radio (r) affects the performance of our model. We first conducted experiments with different window sizes: $w = [50, 100, 150, 200, 250, 300]$. Fig. 6 shows our model's performance in terms of F1 and AUC using varying window sizes. In most cases, our model achieves highest F1 and AUC scores when widow size is 200. Our model tends to yield lower F1 scores when using smaller window sizes. This is because our model may struggle to capture correlations between different time sequences when the input sequence contains limited information. Furthermore, if the window size is too small, there is a significant risk of the masked noise from data augmentation overshadowing the true sequence information. This makes it difficult for our model to learn the distribution of normal data, thereby affecting the accuracy of reconstructing normal data. However, if the window size is too large, the reconstruction errors for normal points within windows containing long abnormal segments may significantly increase. This may result in misjudgments of anomalies when computing anomaly scores.

Fig. 7 shows our model's performance in terms of F1 and AUC using varying mask radios. We conducted experiments with various mask radios: $r = [0.01, 0.05, 0.1, 0.2, 0.3]$. In most cases, our model achieves highest F1 and AUC scores when mask radio is 0.1. When the mask radio is extremely small, our model achieves low F1 and AUC scores. It is possible that when applying a small mask ratio to the MTS, the masked regions become excessively similar, leading to an incomplete representation. However, if the mask radio is excessively large, it leads to an overwhelming proportion of masked data. This excessive masking makes it challenging to bring positive sample pairs closer together.

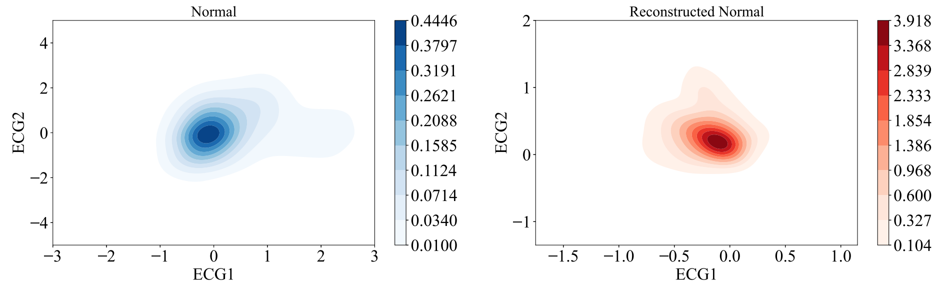

Fig. 8. Distribution of the normal data. The figure on the left illustrates the distribution of the original normal data, while the right illustrates the distribution of the reconstructed normal data.

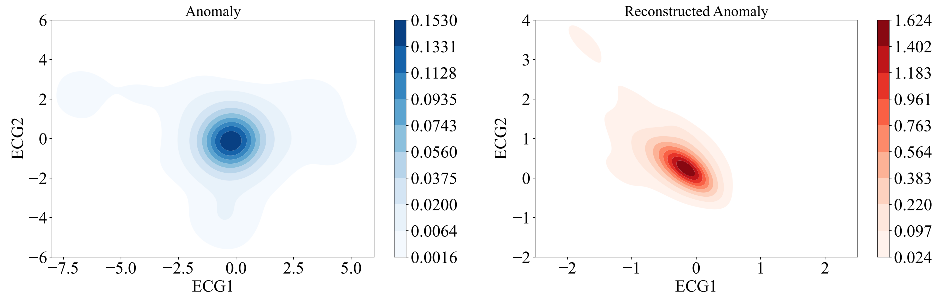

Fig. 9. The distribution of the abnormal data. The figure on the left illustrates the distribution of the original anomaly data, while the right illustrates the distribution of the reconstructed anomaly data.

5.5. Case study

In this section, we selected the MBA dataset, which consists of only two dimensions, to analyze the anomalies from both macro and micro perspectives. In this dataset, many of signals contain high-frequency noise and baseline drift. In addition, due to the patients' movements, there may be some disturbances that appear to be abnormal in certain leads, which are not truly abnormal. Consequently, this poses challenges in anomaly detection in the field of engineering informatics.

In our experiments, we selected 3000 consecutive points from the MBA dataset for better visualization. Each record contains two ECG (Electrocardiogram) signals: ECG1 and ECG2. These two signals represent two distinct electrocardiogram leads recorded from the same patient, with each lead taken from a different angle or position.

First, the distributions of normal and anomaly data are visualized in Figs. 8 and 9. The left-hand side of Fig. 8 illustrates the distribution of normal data, while the right-hand side displays the distribution of reconstructed normal data. The bar on the right side of each figure represents the cumulative distribution. Meanwhile, the left-hand side of Fig. 9 displays the original distribution of anomaly data, and the right-hand side displays the distribution of reconstructed anomaly data. We can observe that the distributions of the reconstructed normal data and original normal data are more similar compared to the distributions of the abnormal data and the reconstructed abnormal data. The results validate that our model exhibits the ability to effectively distinguish anomalies by analyzing their corresponding anomaly scores.

Then, we illustrate the true anomalies and detected anomalies in each dimension based on anomaly scores in Fig. 10. We can observe that numerous extreme points tend to cluster in the central region of the actual anomaly segment. Notably, our model is still able to accurately detect truly anomalies in each dimension. The reason for this may be due to the use of data augmentation method that employs geometrically distributed masks for each dimension. This approach enhances the stability of our model to accommodate the variations present in each dimension. Thus, our model, further trained within a generative adversarial framework, demonstrates an enhanced ability to distinguish normal and abnormal patterns.

6. Conclusion

In this paper, we propose a novel approach for MTS anomaly detection aimed at mitigating the overfitting challenge. Specifically, we use data augmentation techniques with a geometrically distributed mask to expand our training data. Then, we train a transformer-based autoencoder within the framework of the generative adversarial network to capture the underlying normal distribution. In addition, we incorporate a contrastive loss into our discriminator to effectively tune the GAN and ensure good generalization. Extensive comparison experiments demonstrate the superiority of our method over state-of-the-art approaches. In particular, our method is capable of detecting anomalies within long sequences. However, there are multiple facets of our approach that could benefit from further refinement. Especially, our method is primarily designed for regularly sampled multivariate time series data. Given that irregularly sampled data and data with substantial missing values could disrupt the consistent temporal patterns in the data, our proposed model may struggle to capture and model these inherent temporal dependencies. Moreover, the performance of our model is upon several hyperparameters. Identifying the optimal set can be labor-intensive. While our approach has shown promising results, it might demand significant computational resources, making it challenging for real-time processing on devices with limited capabilities.

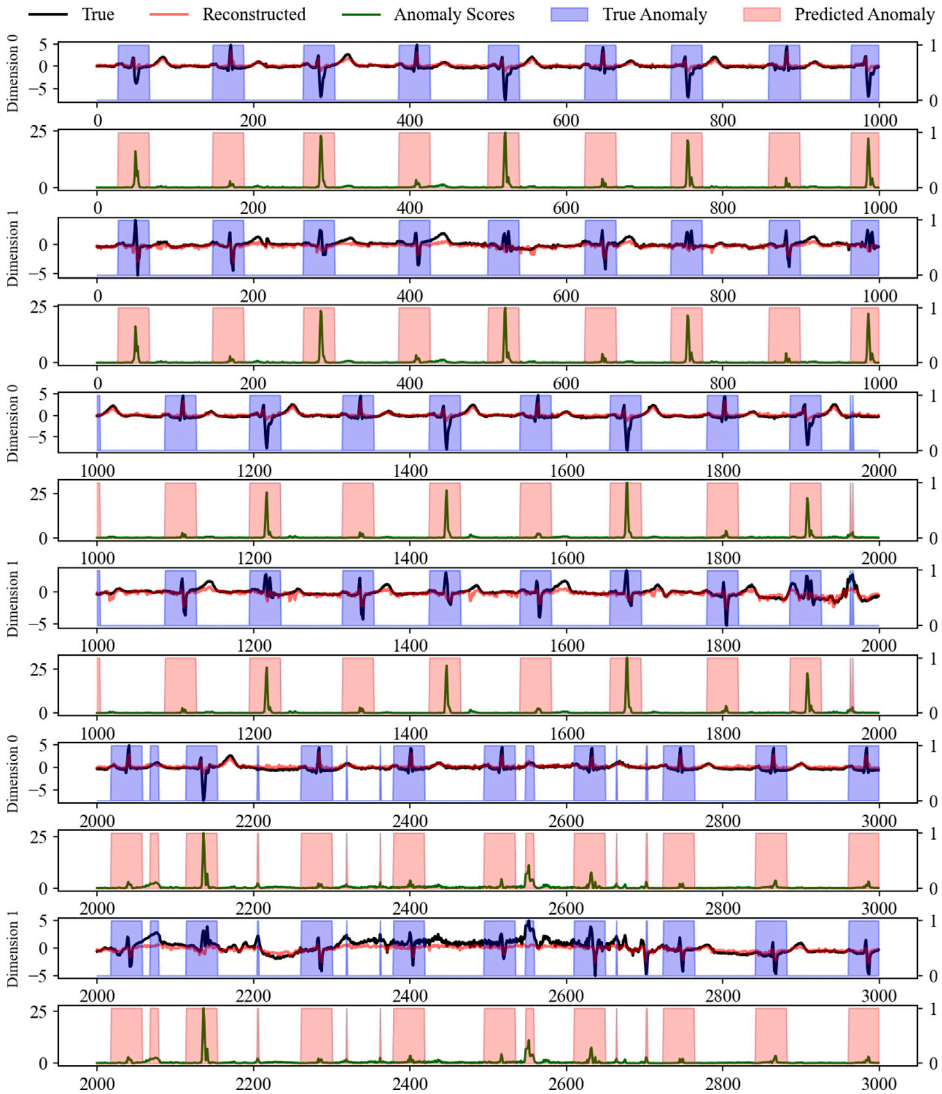

Fig. 10. A comparison between detected and truth anomalies for the MBA dataset. The figure visualizes 3,000 data points collected from the MBA dataset. The black lines represent the original data, while the red lines represent the reconstructed data. The green line corresponds to the anomaly scores obtained from our model. The blue blocks indicate the true anomalies, while the red blocks represent the predicted anomalies based on the anomaly scores. There is a significant overlap between the detected anomalies and the true anomalies, indicating the strong consistency of our anomaly detection method with the actual anomalies in the data. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CRediT authorship contribution statement

Jiawei Miao: Conceptualization, Investigation, Software, Writing – original draft. **Haicheng Tao:** Conceptualization, Methodology, Writing – review & editing, Supervision, Funding acquisition, Writing – original draft. **Haoran Xie:** Formal analysis, Supervision, Writing – review & editing. **Jianshan Sun:** Formal analysis, Supervision, Writing – review & editing. **Jie Cao:** Supervision, Validation, Funding acquisition.

Declaration of competing interest

We declare that we have no competing interests.

Data availability

Data will be made available on request.

Acknowledgments

This research is partially supported by the Key Program of National Natural Science Foundation of China under grant 92046026, and National Natural Science Foundation of China under Grant No. 72172057, and the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province under grant 21KJB520034.

References

Abbasimehr, H., Shabani, M., & Yousefi, M. (2020). An optimized model using LSTM network for demand forecasting. *Computers & Industrial Engineering*, *143*, Article 106435.

Akcay, S., Abarghouei, A. A., & Breckon, T. P. (2018). GANomaly: Semi-supervised anomaly detection via adversarial training. In C. V. Jawahar, H. Li, G. Mori, & K. Schindler (Eds.), *Lecture notes in computer science*: *vol. 11363*, *Computer vision - ACCV 2018 - 14th Asian conference on computer vision, Perth, Australia, December 2-6, 2018, revised selected papers, part III* (pp. 622–637). Springer.

Alamr, A., & Artoli, A. M. (2023). Unsupervised transformer-based anomaly detection in ECG signals. *Algorithms*, *16*(3), 152.

Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2020). USAD: UnSupervised anomaly detection on multivariate time series. In R. Gupta, Y. Liu, J. Tang, & B. A. Prakash (Eds.), *KDD '20: The 26th ACM SIGKDD conference on knowledge discovery and data mining, virtual event, CA, USA, August 23-27, 2020* (pp. 3395–3404). ACM.

Breunig, M. M., Kriegel, H., Ng, R. T., & Sander, J. (2000). LOF: identifying density-based local outliers. In W. Chen, J. F. Naughton, & P. A. Bernstein (Eds.), *Proceedings of the 2000 ACM SIGMOD international conference on management of data, May 16-18, 2000, Dallas, Texas, USA* (pp. 93–104). ACM.

Canizo, M., Triguero, I., Conde, A., & Onieva, E. (2019). Multi-head CNN-RNN for multi-time series anomaly detection: An industrial case study. *Neurocomputing*, *363*, 246–260.

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., & Dubnov, S. (2022). HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022 - 2022 IEEE international conference on acoustics, speech and signal processing* (pp. 646–650). http://dx.doi.org/10.1109/ICASSP43922.2022.9746312.

Chen, Y., Zhou, L., Bouguila, N., Wang, C., Chen, Y., & Du, J. (2021). BLOCK-DBSCAN: Fast clustering for large scale data. *Pattern Recognition*, *109*, Article 107624.

Cheng, D., Wang, X., Zhang, Y., & Zhang, L. (2022). Graph neural network for fraud detection via spatial-temporal attention. *IEEE Transactions on Knowledge and Data Engineering*, *34*(8), 3800–3813. http://dx.doi.org/10.1109/TKDE.2020.3025588.

Cheng, D., Yang, F., Xiang, S., & Liu, J. (2022). Financial time series forecasting with multi-modality graph neural network. *Pattern Recognition*, *121*, Article 108218.

Chopra, S., Hadsell, R., & LeCun, Y. (2005). Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition, vol. 1* (pp. 539–546). IEEE.

Deng, A., & Hooi, B. (2021). Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 5* (pp. 4027–4035).

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics.

Du, B., Sun, X., Ye, J., Cheng, K., Wang, J., & Sun, L. (2021). GAN-based anomaly detection for multivariate time series using polluted training set. *IEEE Transactions on Knowledge and Data Engineering*, 1. http://dx.doi.org/10.1109/TKDE.2021.3128667.

Duan, Z., Xu, H., Wang, Y., Huang, Y., Ren, A., Xu, Z., et al. (2022). Multivariate time-series classification with hierarchical variational graph pooling. *Neural Networks*, *154*, 481–490.

Geiger, A., Liu, D., Alnegheimish, S., Cuesta-Infante, A., & Veeramachaneni, K. (2020). Tadgan: Time series anomaly detection using generative adversarial networks. In X. Wu, C. Jermaine, L. Xiong, X. Hu, O. Kotevska, S. Lu, & et al. (Eds.), *2020 IEEE international conference on big data (IEEE bigData 2020), Atlanta, GA, USA, December 10-13, 2020* (pp. 33–43). IEEE.

Goh, J., Adepu, S., Junejo, K. N., & Mathur, A. (2016). A dataset to support research in the design of secure water treatment systems. In G. M. Havârneanu, R. Setola, H. Nassopoulos, & S. D. Wolthusen (Eds.), *Lecture Notes in Computer Science*: *10242*, *Critical information infrastructures security - 11th International conference, CRITIS 2016, Paris, France, October 10-12, 2016, revised selected papers* (pp. 88–99). Springer.

Han, X., Chen, X., & Liu, L.-P. (2020). GAN ensemble for anomaly detection. arXiv preprint arXiv:2012.07988.

Han, R., Li, P., & Shi, Z. (2022). Implementation strategy of predictive maintenance in nuclear power plant. In *2022 Prognostics and health management conference* (pp. 143–146). http://dx.doi.org/10.1109/PHM2022-London52454.2022.00033.

Hao, S., Wang, Z., Alexander, A. D., Yuan, J., & Zhang, W. (2023). MICOS: Mixed supervised contrastive learning for multivariate time series classification. *Knowledge-Based Systems*, *260*, Article 110158.

Huang, Y. (2023). ViT-r50 GAN: Vision transformers hybrid model based generative adversarial networks for image generation. In *2023 3rd international conference on consumer electronics and computer engineering* (pp. 590–593). http://dx.doi.org/10.1109/ICCECE58074.2023.10135253.

Hundman, K., Constantinou, V., Laporte, C., Colwell, I., & Söderström, T. (2018). Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In Y. Guo, & F. Farooq (Eds.), *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 387–395). ACM.

Jiang, J., Wu, L., Zhao, H., Zhu, H., & Zhang, W. (2023). Forecasting movements of stock time series based on hidden state guided deep learning approach. *Information Processing & Management*, *60*(3), Article 103328.

Khreich, W., Khosravifar, B., Hamou-Lhadj, A., & Talhi, C. (2017). An anomaly detection system based on variable N-gram features and one-class SVM. *Information and Software Technology*, *91*, 186–197.

Kitaev, N., Kaiser, L., & Levskaya, A. (2020). Reformer: The efficient transformer. In *8th International conference on learning representations*. OpenReview.net.

Lee, Y., Byun, Y., & Baek, J. (2023). Time series anomaly detection using contrastive learning based one-class classification. In *International conference on artificial intelligence in information and communication* (pp. 330–335). IEEE.

Li, D., Chen, D., Jin, B., Shi, L., Goh, J., & Ng, S.-K. (2019). MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *International conference on artificial neural networks* (pp. 703–716). Springer.

Ludeña-Choez, J., Zevallos, J. J. C., & Mayhua-López, E. (2019). Sensor nodes fault detection for agricultural wireless sensor networks based on NMF. *Computers and Electronics in Agriculture*, *161*, 214–224.

Lv, J., Wang, Y., & Chen, S. (2023). Adaptive multivariate time-series anomaly detection. *Information Processing & Management*, *60*(4), Article 103383.

Maciąg, P. S., Kryszkiewicz, M., Bembenik, R., Lobo, J. L., & Del Ser, J. (2021). Unsupervised anomaly detection in stream data with online evolving spiking neural networks. *Neural Networks*, *139*, 118–139.

Nakamura, T., Imamura, M., Mercer, R., & Keogh, E. (2020). Merlin: Parameter-free discovery of arbitrary length anomalies in massive time series archives. In *2020 IEEE international conference on data mining* (pp. 1190–1195). IEEE.

Nguyen, H. D., Tran, K. P., Thomassey, S., & Hamad, M. (2021). Forecasting and anomaly detection approaches using LSTM and LSTM autoencoder techniques with the applications in supply chain management. *International Journal of Information Management*, *57*, Article 102282.

Pöppelbaum, J., Chadha, G. S., & Schwung, A. (2022). Contrastive learning based self-supervised time-series analysis. *Applied Soft Computing*, *117*, Article 108397.

Shao, M., Lin, Y., Peng, Q., Zhao, J., Pei, Z., & Sun, Y. (2023). Learning graph deep autoencoder for anomaly detection in multi-attributed networks. *Knowledge-Based Systems*, *260*, Article 110084.

Sohn, K. (2016). Improved deep metric learning with multi-class N-pair loss objective. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems, vol. 29*. Curran Associates, Inc..

Tao, H., Miao, J., Zhao, L., Zhang, Z., Feng, S., Wang, S., et al. (2023). HAN-CAD: hierarchical attention network for context anomaly detection in multivariate time series. *World Wide Web*, 1–16. http://dx.doi.org/10.1007/s11280-023-01171-1.

Thanthawy Sukanda, A. J., & Adytia, D. (2022). Wave forecast using bidirectional GRU and GRU method case study in pangandaran, Indonesia. In *2022 International conference on data science and its applications* (pp. 278–282). http://dx.doi.org/10.1109/ICoDSA55874.2022.9862832.

Tuli, S., Casale, G., & Jennings, N. R. (2022). TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, *15*(6), 1201–1214.

Wang, Y., Li, J., Pei, Y., Ma, Z., Jia, Y., & Wei, Y.-C. (2022). An adaptive high-voltage direct current detection algorithm using cognitive wavelet transform. *Information Processing & Management*, *59*(2), Article 102867. http://dx.doi.org/10.1016/j.ipm.2022.102867.

Woo, G., Liu, C., Sahoo, D., Kumar, A., & Hoi, S. C. H. (2022). Cost: Contrastive learning of disentangled seasonal-trend representations for time series forecasting. In *The tenth international conference on learning representations*. OpenReview.net.

Xi, L., Liang, C., Liu, H., & Li, A. (2023). Unsupervised dimension-contribution-aware embeddings transformation for anomaly detection. *Knowledge-Based Systems*, *262*, Article 110209.

Xiao, B., Zhang, Y., Chen, Y., & Yin, X. (2021). A semi-supervised learning detection method for vision-based monitoring of construction sites by integrating teacher-student networks and data augmentation. *Advanced Engineering Informatics*, *50*, Article 101372.

Xu, H., Chen, W., Zhao, N., Li, Z., Bu, J., Li, Z., et al. (2018). Unsupervised anomaly detection via variational auto-encoder for seasonal KPIs in web applications. In P. Champin, F. Gandon, M. Lalmas, P. G. Ipeirotis (Eds.), *Proceedings of the 2018 World Wide Web conference on World Wide Web* (pp. 187–196). ACM.

Xu, J., Wu, H., Wang, J., & Long, M. (2022). Anomaly transformer: Time series anomaly detection with association discrepancy. In *The tenth international conference on learning representations*. OpenReview.net.

Yang, X., Zhang, Z., & Cui, R. (2022). TimeCLR: A self-supervised contrastive learning framework for univariate time series representation. *Knowledge-Based Systems*, *245*, Article 108606.

Zhang, L., & Liu, L. (2022). Data anomaly detection based on isolation forest algorithm. In *2022 International conference on computation, big-data and engineering* (pp. 87–89). http://dx.doi.org/10.1109/ICCBE56101.2022.9888169.

Zhou, B., Liu, S., Hooi, B., Cheng, X., & Ye, J. (2019). BeatGAN: Anomalous rhythm detection using adversarially generated time series. In S. Kraus (Ed.), *Proceedings of the twenty-eighth international joint conference on artificial intelligence* (pp. 4433–4439). ijcai.org.

Zong, B., Song, Q., Min, M. R., Cheng, W., Lumezanu, C., Cho, D., et al. (2018). Deep autoencoding Gaussian mixture model for unsupervised anomaly detection. In *6th International conference on learning representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, conference track proceedings*. OpenReview.net.