

# Unsupervised Time Series Anomaly Prediction with Importance-based Generative Contrastive Learning

Kai Zhao  
Aalborg University  
Aalborg, Denmark  
kaiz@cs.aau.dk

Zhihao Zhuang  
East China Normal University  
Shanghai, China  
zhuangzhihao@stu.ecnu.edu.cn

Chenjuan Guo\*  
East China Normal University  
Shanghai, China  
cjguo@dase.ecnu.edu.cn

Hao Miao  
Aalborg University  
Aalborg, Denmark  
haom@cs.aau.dk

Christian S. Jensen  
Aalborg University  
Aalborg, Denmark  
csj@cs.aau.dk

Yun Yao Cheng  
Aalborg University  
Aalborg, Denmark  
yunyao@cs.aau.dk

Bin Yang  
East China Normal University  
Shanghai, China  
byang@dase.ecnu.edu.cn

## Abstract

We study the problem of time series anomaly prediction, which is relevant to a range of real-world applications. Existing anomaly prediction methods rely on labeled training data for achieving acceptable accuracy. However, such data may be difficult to obtain; and in real-time deployments, anomalies can occur that were not seen in labeled data, thus making them difficult to predict. We provide a theoretical analysis and propose an Importance-based Generative Contrastive Learning method (IGCL) for unsupervised anomaly prediction. IGCL employs a controlled diffusion module to produce anomaly precursor patterns. Next, IGCL learns contextual representations to extract temporal dependencies from pairs of normal time series and anomaly precursors. IGCL is then able to predict anomalies by identifying anomaly precursors that will evolve into future anomalies. To address challenges caused by potentially complex precursor combinations involving multiple variables, we propose a memory bank with importance scores that stores representative samples adaptively and generates more complex anomaly precursors. Extensive experiments on nine benchmark datasets offer evidence that the proposed method is able to outperform state-of-the-art baselines.

## CCS Concepts

• Computing methodologies → Neural networks.

## Keywords

Time series, Anomaly prediction, Generative contrastive learning

\*Corresponding Author



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

KDD '25, Toronto, ON, Canada

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1454-2/2025/08

<https://doi.org/10.1145/3711896.3737174>

## ACM Reference Format:

Kai Zhao, Zhihao Zhuang, Chenjuan Guo, Hao Miao, Christian S. Jensen, Yun Yao Cheng, and Bin Yang. 2025. Unsupervised Time Series Anomaly Prediction with Importance-based Generative Contrastive Learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25)*, August 3–7, 2025, Toronto, ON, Canada. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3711896.3737174>

## KDD Availability Link:

The source code of this paper has been made publicly available at <https://doi.org/10.5281/zenodo.15561219>.

## 1 Introduction

Much data collected by real-world applications can be modeled as time-dependent observations that form multi-variable time series [10, 45, 49, 51, 53]. For example, computing systems record runtime indicators [1, 38] and environmental systems record water quality [11, 29]. Based on the current time series data, anomaly prediction makes real-time judgments on whether anomaly signals start to occur that indicate that severe anomalies will occur after the current time [4, 7, 17, 21], as exemplified in Figure 1(a). Anomaly prediction plays an important role in scenarios that preserve some form of safety and prioritize losses caused by anomalies [2, 3, 32, 47], e.g., in health care [7] or drinking-water systems [21]. For example, anomaly prediction can help address impending pollution or faults in advance [4, 17] or can warn of extreme environmental conditions [32].

Only a few classification-based studies [2, 3, 32, 47] exist related to the time series anomaly prediction. Existing methods predict future anomalies by determining whether current observations contain anomaly signals that are the start of deviations from normal behaviors [4, 7, 17]. Such early signals are called the anomaly precursors [21]. However, to obtain accurate results, existing methods heavily rely on supervised learning, where anomaly precursors are given during training.

Yet, in many real-world applications, there is insufficient labeled training data to enable supervised learning because manual labeling

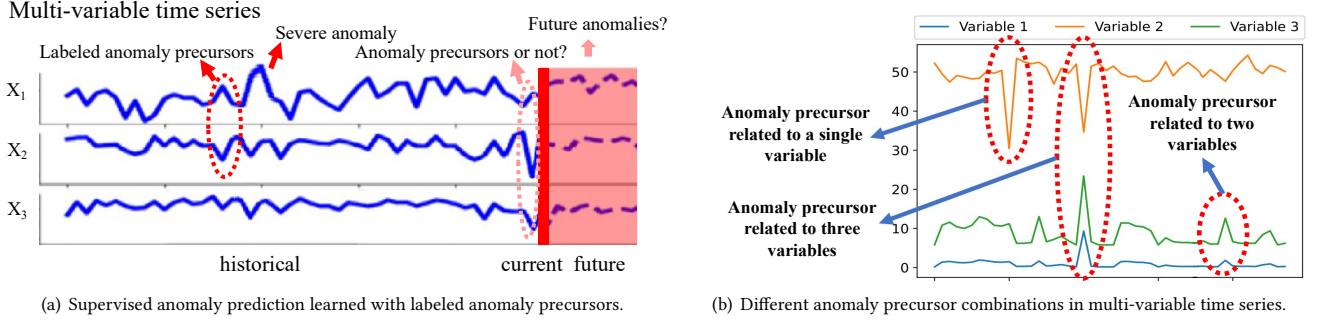


Figure 1: Our motivation.

is costly [35]. Further, unexpected anomalies [16, 43] that did not appear prior to occurring during real-time deployment, and thus were not considered during training and make supervised methods fail in practice [55]. Therefore, unsupervised time series anomaly prediction is of high interest. However, this has not been studied in depth nor analyzed theoretically due to its challenging nature, as outlined next.

**Challenge 1:** In the setting of unsupervised anomaly prediction, there is no labeled anomaly precursor data to enable models to learn temporal dependencies. Although PAD [21] proposes to use unsupervised anomaly detection methods [6, 24, 30] with a specific threshold to identify the precursors, their performance is limited when compared to supervised learning methods [17]. Without labeled data, existing methods cannot learn temporal dependency features from pairs of a successive normal sub-sequence and precursor, where the data just begins to deviate from normal [7], which is crucial to enable anomaly prediction [4, 32]. As a result, existing unsupervised methods do not perform well in anomaly prediction.

**Challenge 2:** We are facing potentially very complex anomaly precursor combinations in multi-variable time series, which cause very high time and space complexity. As shown in Figure 1(b), for any  $n$ -variable time series, there can be as many as  $O(2^n)$  potential anomaly precursor combinations [17], where different anomaly precursors may appear in each individual variable or in variable combinations [52, 54]. It is very challenging to learn temporal dependencies comprehensively from this many anomaly precursor combinations.

We propose a novel anomaly prediction method called IGCL, which is short for Importance-based Generative Contrastive Learning for unsupervised time series anomaly prediction. To address Challenge 1, we employ a generative contrastive learning architecture. We propose an anomaly precursor pattern generation module that employs a diffusion-based transformation and variance regularization, to generate different kinds of potential precursor patterns labeled as negative data using Gaussian noise. We also propose an overlapping window-based temporal convolutional network (TCN) architecture to efficiently extract temporal dependencies and learn positive and negative contextual representations for contrastive learning. More specifically, the model distinguishes normal sub-sequences from anomaly precursors, where pairs of successive normal sub-sequences are positive samples and pairs of a normal

sub-sequence and an anomaly precursor, which contain temporal changes from normal to abnormal, are negative samples.

To address Challenge 2, we propose a memory bank with importance-based scores to adaptively store representative anomaly precursors, which helps generate more complex anomaly precursor combinations related to different variables. First, anomaly precursor patterns can be generated efficiently as only one variable is considered during each iteration. Second, the memory bank, with its fixed size  $K$  that could be far smaller than  $O(2^n)$ , is initialized with generated anomaly precursors related to one variable. Then, other one-variable anomaly precursor patterns are generated iteratively and are injected into the previously generated anomaly precursors stored in the memory bank, thereby accumulating complex anomaly precursor combinations involving more variables. The memory bank only stores representative anomaly precursors and drops anomaly precursors that are already distinguished. We summarize our contributions as follows.

- We propose a generative contrastive learning architecture with a diffusion-based transformation and variance regularization to generate different precursor patterns and learn temporal dependencies to distinguish between normal and anomaly precursors.
- We propose a memory bank with importance scores to store representative negative samples, which helps to generate more complex negative samples.
- Experiments on nine benchmarks from different domains show our method outperforms the state-of-the-art baselines.

## 2 PRELIMINARIES

### 2.1 Definitions

We first formalize the unsupervised time series anomaly prediction problem. Frequently used notation is summarized in Table 1.

**A Multi-variable Time Series** is represented as  $\mathbf{X} = \langle \mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T \rangle \in \mathbb{R}^{N \times T}$ , where  $N$  is the number of variables collected at each timestamp,  $T$  is the total number of historical timestamps,  $\mathbf{x}_t \in \mathbb{R}^N$  denotes the  $N$ -dimensional observations at timestamp  $t$ . We use  $\mathbf{X}_{t+1:t+f} \in \mathbb{R}^{N \times f}$  to indicate the data during the time window  $[t+1, t+f]$ , use  $\mathbf{X}_{t+1:t+f}^i \in \mathbb{R}^f$  to indicate the data of the  $i$ -th variable, and use  $\mathbf{X}_t^i \in \mathbb{R}^1$  to indicate the data of the  $i$ -th variable at timestamp  $t$ . We also use  $T$  to denote the current timestamp.

**Table 1: Notation**

Notation	Explanation
$\mathbf{x}_t$	The observed data at timestamp $t$ from the $N$ -variable time series $\mathbf{X}$
$\mathbf{X}_{t+1:t+f}$	The sub-sequence of the time series $\mathbf{X}$ from time $t + 1$ to $t + f$ , with sequence length of $f$
$\mathbf{X}_{t-h:t}^i$	The data of the $i$ -th variable starting from $t - h$ to $t$
$\mathbb{X}_{T,h}^b$	The set of most recent $b$ sub-sequences with length $h + 1$ before the current timestamp $T$ $\mathbb{X}_{T,h}^b = \{\mathbf{X}_{T-h:T}, \dots, \mathbf{X}_{T-b+1-h:T-b+1}\}$
$N(\mu, \sigma^2)$	The multi-variable Gaussian distribution with mean vector $\mu$ and covariance matrix $\sigma^2$
$\mathbf{R}_{t-h:t}^i$	The anomaly precursor patterns for the $i$ -th variable
$S$	The max steps for diffusion-based transformation
$\mathbf{v}_t$	The feature vector extracted from time series at timestamp $t$
$z_t^+$	The representations from $(\mathbf{X}_{t-2h-1:t-h-1}, \mathbf{X}_{t-h:t})$ to estimate $\mathcal{N} = P(\cdot   \mathbf{H}_{t-h}, \mathbf{H}_t \text{ is normal})$
$z_t^-$	The representations from $(\mathbf{X}_{t-2h-1:t-h-1}, \mathbf{X}_{t-h:t}^-)$ to estimate $\mathcal{A} = P(\cdot   \mathbf{H}_{t-h}, \mathbf{H}_t \text{ is an anomaly precursor})$
$K$	The max size of the memory bank

**A Sub-sequence**  $\mathbf{X}_{t_1:t_2}$  is a continuous subset of  $\mathbf{X}$  from  $t_1$  to  $t_2$ . The set of  $b$  most recent sub-sequences before the current timestamp  $T$ , each with sequence length of  $h + 1$ , is denoted as  $\mathbb{X}_{T,h}^b = \{\mathbf{X}_{T-h:T}, \mathbf{X}_{T-1-h:T-1}, \dots, \mathbf{X}_{T-b+1-h:T-b+1}\}$ .

**An Anomaly Precursor** is a sub-sequence  $\mathbf{X}_{T-h:T}$  containing early signals that just start deviating from normal and followed by future anomalies in  $\mathbf{X}_{T+1:T+f}$  [4, 17, 21], where  $h$  and  $f$  are hyper-parameters for the look-back and look-forward windows.

**Unsupervised time series anomaly prediction.** At any current timestamp  $T$ , given the historical observations  $\mathbf{X}$ , where we use a set of  $b$  most recent sub-sequences  $\mathbb{X}_{T,h}^b = \{\mathbf{X}_{T-h:T}, \mathbf{X}_{T-1-h:T-1}, \dots, \mathbf{X}_{T-b+1-h:T-b+1}\}$ , the goal is to determine a real-time anomaly score  $\hat{p}_T$  to indicate whether there are anomaly precursors currently and more anomalies in  $\mathbf{X}_{T+1:T+f}$ . The transformation of anomaly scores into binary labels is done by applying a threshold  $\delta$  [21], i.e., if  $\hat{p}_T < \delta$ , the future sub-sequence  $\mathbf{X}_{T+1:T+f}$  is normal; and if  $\hat{p}_T \geq \delta$ , the future sub-sequence  $\mathbf{X}_{T+1:T+f}$  is abnormal. Thus, we formulate the time series anomaly prediction problem as that of finding a model with a mapping function  $\mathcal{F}$  and  $\hat{p}_T = \mathcal{F}_\theta(\mathbb{X}_{T,h}^b)$ , where  $\theta$  is the parameters of the model  $\mathcal{F}$ . For the unsupervised time series anomaly prediction problem, we do not have the true score  $p_T$  as there is no labeled data.

## 2.2 Theoretical Analysis

We analyze unsupervised time series anomaly prediction using Maximum Mean Discrepancy (MMD) theory [20]. We utilize Markov Chains [28], widely used in time series analytics [13], to model sequence-based probabilities. We use random variables  $\mathbf{H}_{t-h-1}$  and  $\mathbf{H}_t$  to denote what observations the sub-sequences  $\mathbf{X}_{t-2h-1:t-h-1}$  and  $\mathbf{X}_{t-h:t}$  may be observed during two successive time-windows  $[t - 2h - 1, t - h - 1]$  and  $[t - h, t]$ , respectively. Then, we define

$$\mathcal{N} = P(\cdot | \mathbf{H}_{t-h-1} = \mathbf{X}_{t-2h-1:t-h-1}, \mathbf{H}_t \text{ is normal})$$

to denote the Markov conditional distribution of the normal sub-sequence observed during time-window  $[t - h, t]$ , given that its

previous observations during time-window  $[t - 2h - 1, t - h - 1]$  are  $\mathbf{X}_{t-2h-1:t-h-1}$ . Similarly, we define

$$\mathcal{A} = P(\cdot | \mathbf{H}_{t-h-1} = \mathbf{X}_{t-2h-1:t-h-1}, \mathbf{H}_t \text{ is anomaly precursor})$$

to denote the Markov conditional distribution of the anomaly precursor observed during time-window  $[t - h, t]$ . We present the theorem here and its proof in Appendix A.

**Theorem:** Under the unsupervised setting, the future anomaly is unpredictable if we cannot identify its anomaly precursor in the current observation that follows the same distribution as the normal previous sub-sequences do. The future anomaly is predictable if we can identify its anomaly precursor that follows a distribution different from the normal previous sub-sequences.  $\square$

Thus, our model aims to distinguish normal distribution  $\mathcal{N} = P(\cdot | \mathbf{H}_{t-h-1}, \mathbf{H}_t \text{ is normal})$  from any potential different distributions  $\mathcal{A} = P(\cdot | \mathbf{H}_{t-h-1}, \mathbf{H}_t \text{ is anomaly precursor})$ .

## 3 Methodology

We present the overall architecture in Figure 2, which consists of anomaly precursor pattern generation, positive and negative representations learning, contrastive loss, and the memory bank.

For any timestamp  $t$ , the model takes as input the set of  $b$  available historical sub-sequences  $\mathbb{X}_{t,h}^b = \{\mathbf{X}_{t-h:t}, \mathbf{X}_{t-1-h:t-1}, \dots, \mathbf{X}_{t-b+1-h:t-b+1}\}$ . We first apply Instance Normalization [40] on the observation of each time series variable, i.e.,  $\mathbf{X}_t^i$ , in the set  $\mathbb{X}_{t,h}^b$ , and output  $\bar{\mathbf{X}}_{t,h}^b$ , to learn stable components [25] from time series.

Embedding layer takes as input the normalized  $\bar{\mathbf{X}}_{t,h}^b$ , maps the observations at each timestamp  $t'$ , i.e.,  $\mathbf{X}_{t'} \in \mathbb{R}^N$ , to a  $d$ -dimensional vectors  $\mathbf{v}_{t'} \in \mathbb{R}^d$ , which extracts the dense feature from normal time series, and outputs a set of feature vectors, i.e.,  $\mathbb{V}^{0,+}$ .

**Anomaly precursor pattern generation** takes as input a randomly generated Gaussian noise  $x \sim N(0, I)$ , and iteratively outputs any potential anomaly precursor pattern, i.e.,  $\mathbf{R}_{t-h:t}^i$ , for a randomly selected  $i$ -th time series variable.

Pattern injection module injects  $\mathbf{R}_{t-h:t}^i$  into the last sub-sequence  $\mathbf{X}_{t-h:t}$  in the selected  $i$ -th time series variable, and outputs a time series anomaly precursor, i.e.,  $\mathbf{X}_{t-h:t}^- = \mathbf{X}_{t-h:t}^i + \mathbf{R}_{t-h:t}^i$ , which is fed into the embedding layer to get precursor feature vectors  $\mathbb{V}^{0,-}$ .

**Positive and negative representations learning** takes as input all possible pairs of successive sub-sequences from  $\mathbb{V}^{0,+}$  and  $\mathbb{V}^{0,-}$ . For example, this module takes as input a pair of normal sub-sequences  $\mathbf{X}_{t-2h-1:t-h-1}$  and  $\mathbf{X}_{t-h:t}$  from  $\mathbb{V}^{0,+}$ , and outputs a positive contextual representation  $z_t^+$  that learns their normal temporal dependencies. This module also takes as input a pair of a normal sub-sequence  $\mathbf{X}_{t-2h-1:t-h-1}$  from  $\mathbb{V}^{0,+}$  and an anomaly precursor  $\mathbf{X}_{t-h:t}^-$  from  $\mathbb{V}^{0,-}$ , and outputs a negative contextual representation  $z_t^-$  that learns their abnormal temporal dependencies.

With **contrastive loss**, the model distinguishes  $\mathcal{N} = P(\cdot | \mathbf{H}_{t-h-1}, \mathbf{H}_t \text{ is normal})$  and  $\mathcal{A} = P(\cdot | \mathbf{H}_{t-h-1}, \mathbf{H}_t \text{ is anomaly precursor})$  by pushing  $z_t^-$  away from  $z_t^+$ . Intuitively, it clusters positive representations, e.g.,  $z_t^+$ , that belong to  $\mathcal{N}$  together and dissociates negative representations, e.g.,  $z_t^-$ , that belong to  $\mathcal{A}$  away.

**The memory bank** stores the representative anomaly precursors. The generated anomaly precursor patterns are iteratively injected into existing anomaly precursors in the memory bank, to accumulate more complex anomaly precursor combinations. The

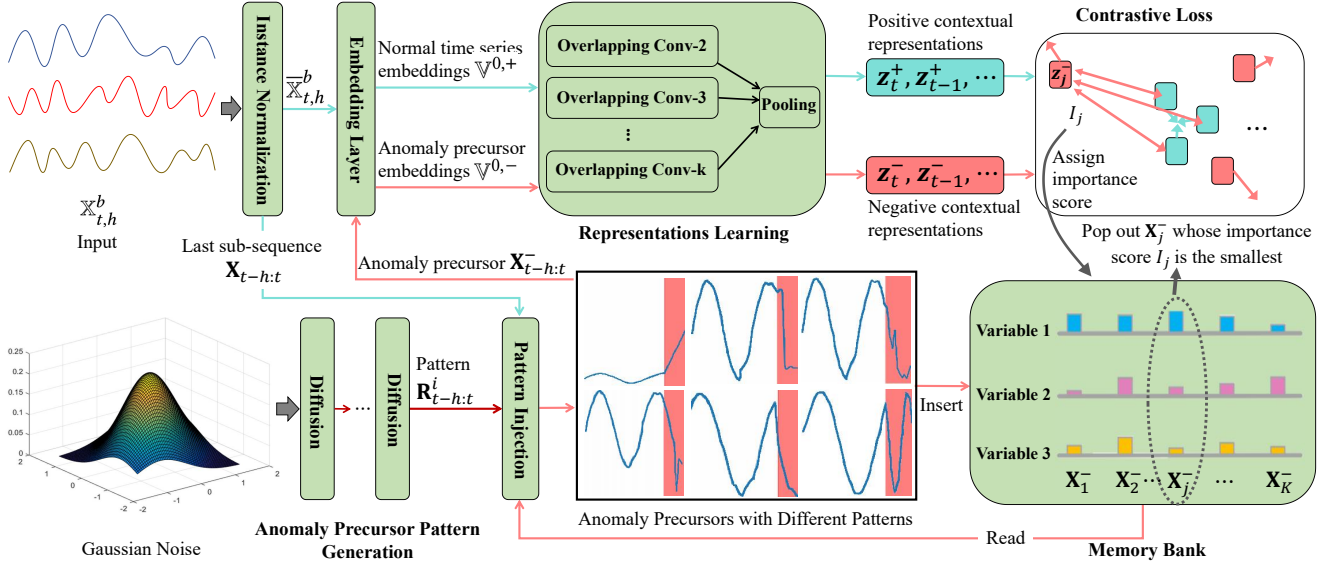


Figure 2: The overall architecture.

importance score  $I_j$  for the anomaly precursor  $X_j^-$  is calculated by its similarity with the normal samples. Finally, the memory bank pops out the anomaly precursor with the smallest importance score, which the model can easily distinguish from normal already.

### 3.1 Anomaly Precursor Pattern Generation

As experimental results show diffusion models outperform other generative models for time series [42, 46], we propose diffusion-based transformation to generate diverse kinds of potential anomaly precursors from a simple Gaussian distribution, with our variance regularization to simulate precursor patterns that start deviating from normal.

The denoising diffusion model consists of two processes, *i.e.*, the pollution process with  $S$  steps and its reverse denoising diffusion process, where  $S$  is the max diffusion steps. Specifically, we use  $x^0 = R_{t-h:t}^i \in \mathbb{R}^h$  to denote any potential anomaly precursor pattern that could be more complicated than Gaussian patterns. The pollution process is shown in Appendix B. Our anomaly precursor pattern generation module is based on denoising diffusion process  $p_\theta(x^{s-1}|x^s)$ , reversing random Gaussian noise  $x^S$  to produce more complicated potential anomaly precursor patterns:

$$x^{s-1} \sim p_\theta(x^{s-1}|x^s) = \frac{1}{\sqrt{1-\beta^s}} \left( x^s - \sqrt{\beta^s} \varepsilon_\theta(x^s) \right), \quad (1)$$

where  $\varepsilon_\theta(x^s) \sim N(\mu_s, \sigma_s^2)$ , and  $\mu_s$  is the mean value parameter and  $\sigma_s$  is the variance parameter to be learned. We use Multi-layer Perceptron (MLP) layers to model with  $\mu^s$  and  $\sigma^s$ :

$$\mu^s, \sigma^s = \text{MLP}(\text{concat}(x^s, s, X_{t-h:t}^i)), \quad \forall s \in \{S, S-1, \dots, 1\},$$

where  $x^S$  is a random Gaussian noise,  $x^s$  is the denoised state after each step, and  $i$  denotes this process aiming to generate anomaly precursor patterns for the  $i$ -th variable. We have  $\sigma^s \varepsilon + \mu^s \sim N(\mu_s, \sigma_s^2)$ , where  $\varepsilon \sim N(0, I)$ , and  $\sigma^s \varepsilon + \mu^s$  is derivable with respect to  $\mu^s$  and  $\sigma^s$  which are the outputs from the MLP layers. Thus,

the final denoising diffusion process is

$$x^{s-1} = \frac{1}{\sqrt{1-\beta^s}} \left( x^s - \sqrt{\beta^s} (\sigma^s \varepsilon + \mu^s) \right), \quad (2)$$

where we randomly sample the Gaussian noise  $x^S$ , and apply the denoising diffusion process step by step to get the anomaly precursor pattern  $R_{t-h:t}^i = x^0$ .

We further propose regularization loss on parameters  $N(\mu_s, \sigma_s^2)$ , where the variance  $\sigma_s^2$  controls the diversity of generated anomaly precursor patterns, and  $\sigma_s^2$  and the mean value  $\mu_s$  controls the degree of deviation from normal. First, the variance  $\sigma_s^2$  should not be too large, as a large variance will produce severe anomalies rather than anomaly precursors with starting deviation from normal. Second,  $\sigma_s^2$  should not be too small, as a small variance will generate less diverse [46] precursor patterns. Therefore, we propose to use the Kullback-Leibler divergence as regularization to make  $N(\mu_s, \sigma_s^2)$  be close to  $N(\mu_s, I)$  where  $\sigma_s^2$  should be close to the unit normal variance  $I$  which is neither too large nor too small [27]. For any  $N(\mu, \sigma^2)$ , we have

$$KL(N(\mu, \sigma^2) \| N(\mu, I)) = \frac{1}{2} \left( -\log \sigma^2 + \sigma^2 - 1 \right). \quad (3)$$

The proof is in Appendix C. Thus, our analytical solution of the regularization loss is:

$$\mathcal{L}_r = \sum_{s=1}^S \frac{1}{2} \left( -\log \sigma_s^2 + \sigma_s^2 - 1 \right) \quad (4)$$

### 3.2 Positive and Negative Representations

We propose an overlapping window-based TCN to learn contextual representations from the all possible pairs of successive subsequences followed by the Markov Chain and extract the normal and abnormal temporal dependencies at once.

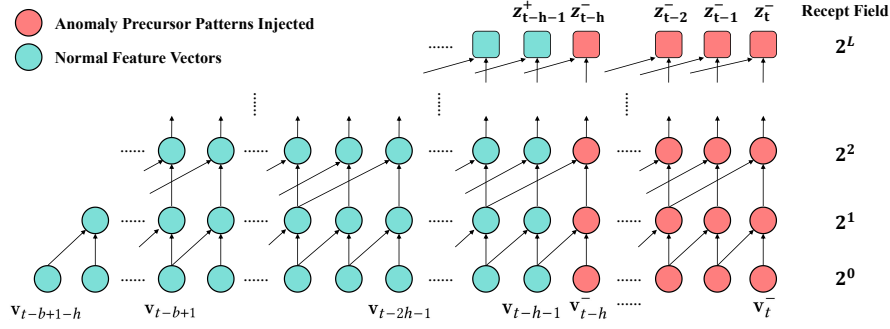


Figure 3: The overlapping window-based temporal convolutional network with a kernel size  $k$  of 2.

To be more specific, our embedding layer firstly maps the normalized time series data, i.e.,  $X_{t'} \in \mathbb{R}^N$  with  $t' \in [t - b + 1 - h, t]$ , equivalent to the set  $\mathbb{X}_{t,h}^b$ , into the  $d$ -dimensional feature vectors  $\mathbf{v}_{t'} \in \mathbb{R}^d$ , which aim to extract the dense features from time series for each timestamp. Further, the input to the embedding layer can also be coupled with other auxiliary attributes, such as the encoding of timestamps. The auxiliary attributes at timestamp  $t'$  are represented as  $\mathbf{a}_{t'} \in \mathbb{R}^F$ , where  $F$  is the total dimensions of the auxiliary attributes. Thus, the embedding layer concatenates the original time series data  $X_{t'}$  and the auxiliary attributes

$$\mathbf{f}_{t'} = \text{concat}(X_{t'}, \mathbf{a}_{t'}) \quad \forall t' \in [t - b + 1 - h, t], \quad (5)$$

as its input, and maps it to the feature vector by:

$$\mathbf{v}_{t'} = \sigma(W_e \mathbf{f}_{t'}), \quad \forall t' \in [t - b + 1 - h, t], \quad (6)$$

where  $\mathbf{v}_{t'} \in \mathbb{R}^d$  is the learned feature vector,  $\sigma$  is an activation function, and  $W_e \in \mathbb{R}^{d \times (N+F)}$  is the learnable neural network parameters which extract the feature from the time series data and the auxiliary attributes at each timestamp. Following the same way, we can also get the feature vector  $\mathbf{v}_{t'}^- \in \mathbb{R}^d$ , where  $t' \in [t - h, t]$ , from the generated anomaly precursor  $\mathbf{X}_{t-h:t}^-$ .

Then, our overlapping window-based TCN uses the feature vectors  $\mathbf{v}_{t'}$  and  $\mathbf{v}_{t'}^-$  to learn the temporal dependencies for all possible pairs of successive sub-sequences at once. The temporal convolution operation slides over timestamps by skipping feature vectors with a certain dilation factor in different layers, as illustrated in Figure 3. More specifically, this module takes as input all the feature vectors, i.e.,  $\mathbf{v}_{t'}$  with  $t' \in [t - b + 1 - h, t]$  and  $\mathbf{v}_{t'}^-$  with  $t' \in [t - h, t]$ :

$$\begin{aligned} \mathbb{V}^{0,+} &= \langle \mathbf{v}_{t-b+1-h}, \dots, \mathbf{v}_{t-h}, \dots, \mathbf{v}_t \rangle, \\ \mathbb{V}^{0,-} &= \langle \mathbf{v}_{t-b+1-h}, \dots, \mathbf{v}_{t-h}^-, \dots, \mathbf{v}_t^- \rangle, \end{aligned} \quad (7)$$

and uses the learnable convolution  $\text{filter} \in \mathbb{R}^k$  with the kernel size of  $k$  to extract features and learn the temporal dependencies by:

$$\begin{aligned} \mathbb{V}^{l,+}(t') &= \sum_{i=0}^{k-1} \text{filter}(i) \mathbb{V}^{l-1,+}(t' - k^{l-1} \times i), \text{ for } 1 \leq l \leq L, \\ \mathbb{V}^{l,-}(t') &= \sum_{i=0}^{k-1} \text{filter}(i) \mathbb{V}^{l-1,-}(t' - k^{l-1} \times i), \text{ for } 1 \leq l \leq L, \end{aligned} \quad (8)$$

where  $k^{l-1}$  is the dilation factor in the layer  $l - 1$ ,  $\mathbb{V}^{l,+}(t')$  and  $\mathbb{V}^{l,-}(t')$  is the learned dependency features in the layer  $l$  for timestamp  $t'$ , and  $L \sim O(\log(h))$  is the max number of TCN layers

whose receptive field is large than the whole window size of the pair of successive sub-sequences. We use different kernel sizes to extract features in parallel and pool them together, to capture different kinds of temporal dependencies with different kernel scales. Finally, it outputs all contextual representations for all pairs of successive sub-sequences at once, e.g.,  $z_t^+ = \mathbb{V}^{L,+}(t)$ ,  $z_{t-1}^+ = \mathbb{V}^{L,+}(t-1)$ ,  $z_t^- = \mathbb{V}^{L,-}(t)$  and  $z_{t-1}^- = \mathbb{V}^{L,-}(t-1)$ . In this way, our total complexity is reduced to  $O((b+h)\log(h))$ , rather than the complexity of all pairs of sub-sequences as  $O((b+h)^2)$ .

### 3.3 The Objective Function

We propose a contrastive strategy to distinguish  $\mathcal{N}$  and  $\mathcal{A}$  by pushing  $z_t^-$  away from  $z_t^+$ . Intuitively, it clusters positive representations, e.g.,  $z_t^+$ , that belong to  $\mathcal{N}$  together and dissociates negative representations, e.g.,  $z_t^-$ , that belong to  $\mathcal{A}$  away. Each anchor  $z_t^+$  is the contextual representation from the pair of the successive normal sub-sequences before timestamp  $t$ , which represents for  $\mathcal{N} = P(\cdot | \mathbf{H}_{t-h}, \mathbf{H}_t \text{ is normal})$ . Thus, we make the anchor  $z_t^+$  similar to the positive representations  $z_{t-j}^+$  with  $1 \leq j \leq P$ , which are from the most nearby successive normal sub-sequences before timestamp  $t$  and they follow the normal data distribution. In order to avoid data leakage and be consistent with the inference stage, for each anchor  $z_t^+$  the positive samples are always from the previous timestamps. Meanwhile, the anchor  $z_t^+$  should be different from the negative representation  $z_t^-$ , which learns the temporal dependencies from normal to the generated anomaly precursors and represents for anomaly precursors distribution  $\mathcal{A}$ . Thus, our contrastive loss function is:

$$\mathcal{L}_c = \sum_{i=t}^{t-h} -\log \frac{\sum_{j=1}^P \exp(\text{Sim}(z_i^+, z_{i-j}^+)/\tau)}{\exp(\text{Sim}(z_i^+, z_i^-)/\tau) + \sum_{j=1}^P \exp(\text{Sim}(z_i^+, z_{i-j}^-)/\tau)}, \quad (9)$$

where  $\tau$  is the temperature parameter to control the softmax strength,  $P$  is the hyper-parameter for the number of positive samples used, and  $\text{Sim}$  is the similarity measurement, e.g., Cosine Similarity, to distinguish whether the contextual representations belong to the same distributions.

Finally, together with the regularization loss, we train the model using Adam gradient descent [26]:

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_r, \quad (10)$$



where  $\lambda$  is to control the strength of our diffusion regularization.

### 3.4 Memory Bank

We propose to use a memory bank to store previously generated anomaly precursors and combine them with the current anomaly precursor pattern  $\mathbf{R}_{t-h:t}^i$ , to accumulate more complex anomaly precursor combinations related to more than one variable.

To be more specific, the memory bank  $M$  has a fixed size of  $K$  and empirically  $K \ll O(2^n)$  in our experiments:

$$\mathbf{M} = \{\mathbf{X}_1^-, \mathbf{X}_2^-, \dots, \mathbf{X}_K^-\}, \quad (11)$$

where each  $\mathbf{X}_j^-$  denotes a stored anomaly precursor. The currently generated anomaly precursor pattern  $\mathbf{R}_{t-h:t}^i$  which is only related to the  $i$ -th variable, is also injected into the anomaly precursors in the memory bank:

$$\mathbf{X}_j^- \leftarrow \mathbf{X}_j^- + \mathbf{R}_{t-h:t}^i, \quad (12)$$

and thus we can get anomaly precursors that involve more than one variable. Then, similar to Equation (7), we can get more negative contextual representations, e.g.,  $z_{t,j}^-$  and  $z_{t-1,j}^-$ , from the anomaly precursor  $\mathbf{X}_j^-$ . The contrastive loss  $\mathcal{L}_c$  is updated to

$$\sum_{i=t}^{t-h} -\log \frac{\sum_{j=1}^P \exp(\text{Sim}(z_i^+, z_{i-j}^+)/\tau)}{\sum_{j=0}^K \exp(\text{Sim}(z_i^+, z_{i,j}^-)/\tau) + \sum_{j=1}^P \exp(\text{Sim}(z_i^+, z_{i-j}^+)/\tau)}, \quad (13)$$

where  $z_{i,0}^- = z_i^-$  is the negative contextual representations from the current anomaly precursor  $\mathbf{X}_{t-h:t}^-$ , and  $z_{i,j}^-$  is the negative contextual representations from the  $j$ -th anomaly precursor  $\mathbf{X}_j^-$  stored in the memory bank.

Finally, the current anomaly precursor  $\mathbf{X}_{t-h:t}^-$  is inserted into the memory bank which will have a size of  $K + 1$ , i.e.,  $\mathbf{X}_0^- = \mathbf{X}_{t-h:t}^-$ . Instead of using a first-in-first-out strategy to maintain the memory bank with a fixed size of  $K$ , we propose to pop out the anomaly precursor based on an importance score. The intuition is that the memory bank should store the hard and important negative samples for further model training, i.e., the anomaly precursors that are not yet distinguished by the model, and should pop out not important samples, i.e., the anomaly precursors that are already dissociated away from normal. The importance score  $I_j$  for the anomaly precursor  $\mathbf{X}_j^-$  is calculated by:

$$I_j = \sum_{i=t}^{t-h} \text{Sim}(z_i^+, z_{i,j}^-). \quad (14)$$

Thus, we pop out the anomaly precursor with the smallest importance score after each iteration of model optimization.

### 3.5 Inference

For each real-time  $T$  in the inference stage, we output the probability score  $\hat{p}_T$  to predict how likely there is an anomaly precursor currently and will be more future anomalies in  $\mathbf{X}_{T+1,T+f}$ :

$$\hat{p}_T = \sum_{j=1}^K \text{Sim}(z_T^+, z_{T,j}^-) - \sum_{j=1}^P \text{Sim}(z_T^+, z_{T-j}^+), \quad (15)$$

where  $K$  is the size of the memory bank with anomaly precursors as negative samples, and  $P$  is the number of positive samples.

## 4 EXPERIMENTS

### 4.1 Experimental Settings

**4.1.1 Datasets.** Following existing works [6, 21, 44], we validate the performance of our method with nine real-world benchmark datasets: SMD [38], PSM [1], MSL [19], SMAP [19], SWAN [29], SWaT [48], GECCO [29], UCR [44] and Credit [55]. More details of the datasets can be seen in Appendix D.1 and Table 2.

**4.1.2 Baselines.** We compare the unsupervised anomaly prediction method PAD [21]. We also modify more state-of-the-art unsupervised anomaly detection methods into anomaly prediction following PAD:

- Density-based: DAGMM [56] and IForest [22].
- Clustering-based: K-means [35], Deep-SVDD [34] and THOC [36].
- Contrastive-based: DCdetector [50] and PAD [21].
- Autoregression and reconstruction-based: LSTM-VAE [23], ATransformer [48], Omni [38], GANomaly [12], CAE-Ensemble [6], PUAD [30] and D3R [42].

We introduce the details of baselines in Section 5. More details of settings are in Appendix D.2.

**4.1.3 Evaluation Metrics.** Following the evaluation methodology in existing works [6, 21, 55], we validated the performance of time series anomaly prediction models with Precision (P), Recall (R), F1-score (F1), and ROC-AUC [33], where larger values indicate higher model performance and ROC-AUC enables evaluation without choosing the threshold. We do not use the point adjustment (PA) metrics, as many recent works have demonstrated that PA can lead to faulty performance evaluations [18, 33, 39, 42, 55]. PA uses true labels to adjust the outputs of models, and it is known that using PA can result in state-of-the-art performance even with random guess [18, 39].

### 4.2 Overall Comparison and Analysis

Table 3 and Table 9 in Appendix D.3 show the overall anomaly prediction performance of different models on all datasets. We randomly repeat each method 3 times and report the average result, where the best F1-score and AUC-ROC are highlighted in bold and significantly outperform the underlined second-best ones.

Key observations are followed. First, IGCL consistently outperforms the state-of-the-art baseline methods on all datasets. It demonstrates that IGCL is able to learn the normal temporal dependencies and identify the abnormal temporal dependencies that change from the normal to the abnormal with our generated anomaly precursors, which finally improves the accuracy of anomaly prediction.

Second, we observe that the contrastive-based, density-based, and clustering-based methods are unsatisfactory on most datasets. This is because they cannot generate anomaly precursors as negative samples to learn the abnormal temporal dependencies that change from normal to abnormal. They only learn with the non-labeled data as positive samples. Thus, their accuracy is limited as anomaly prediction requires learning both normal patterns and abnormal temporal dependencies.

Third, IGCL achieves the best accuracy compared to the autoregression and reconstruction-based methods, as they only learn to reconstruct each normal time series sub-sequence within each time

**Table 2: Details of benchmark datasets.**

Dataset	Domain	$N$	#Training (Unlabeled)	#Validation (Unlabeled)	#Test (Labeled)	Anomaly Rate(%)
SMD	server machine	38	566,724	141,681	708,420	4.16
PSM	application server	25	105,885	26,398	87,841	27.8
MSL	NASA space sensor	55	46,653	11,663	73,729	10.72
SMAP	NASA space sensor	25	108,146	27,036	427,617	13.13
SWAN	space solar weather	38	48,000	12,000	60,000	32.6
SWaT	water treatment	51	396,000	99,000	449,919	12.14
GECCO	water quality	9	55,408	13,852	69,261	1.1
UCR	various domains	1	1,790,679	447,670	6,143,541	0.6
Credit	finance	29	159,491	39,873	85,443	0.172

**Table 3: Overall performance. We report Precision, Recall, and F1-score results here, presented in percentages.**

Dataset	PSM			SMAP			SWAN			SWaT			GECCO			SMD			MSL		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
DAGMM	39.27	36.14	37.64	11.91	22.00	15.51	52.73	39.46	45.14	82.14	61.72	70.48	47.50	27.64	34.95	14.74	15.37	15.05	13.94	17.04	15.33
IForest	39.39	38.82	39.10	11.79	18.67	14.45	53.85	39.07	45.28	75.39	62.22	68.17	42.75	29.89	35.18	15.25	18.28	16.63	13.82	17.11	15.29
K-means	38.89	39.04	38.96	13.05	21.37	16.20	54.04	42.30	47.45	82.13	62.50	70.98	47.94	29.18	36.28	15.77	17.40	16.54	14.23	15.99	15.06
Deep-SVDD	36.63	36.30	36.46	11.77	17.09	13.94	52.39	39.18	44.83	81.29	61.95	70.31	46.09	28.45	35.18	14.99	15.78	15.37	12.94	16.04	14.32
THOC	37.64	38.93	38.27	12.61	21.89	15.99	54.93	40.02	46.30	82.42	62.45	71.06	46.41	29.49	36.09	16.90	16.79	16.84	14.37	16.22	15.24
DCdetector	28.97	12.05	17.02	8.04	11.30	9.40	21.04	10.94	14.40	46.45	36.36	40.79	1.76	3.46	2.33	3.71	9.03	5.26	2.35	3.46	2.80
PAD	31.23	33.05	32.11	11.01	22.37	14.76	50.13	37.61	42.98	74.83	59.09	66.04	44.38	28.17	34.46	11.23	18.76	14.05	14.37	13.22	13.77
ATransformer	30.56	34.64	32.47	11.40	22.80	15.20	52.20	38.75	44.48	75.00	60.50	66.97	43.62	27.92	34.05	10.52	19.42	13.65	14.62	12.42	13.43
LSTM-VAE	39.20	40.31	40.02	11.68	22.65	15.41	54.42	40.88	46.69	66.76	68.36	67.55	46.67	30.92	37.20	14.21	16.42	15.24	16.18	15.28	15.72
Omni	39.45	40.88	40.15	11.90	23.39	15.77	54.83	41.00	46.92	85.73	58.57	69.59	50.55	29.65	37.38	16.22	18.19	17.15	13.66	21.66	16.75
GANomaly	37.02	43.06	39.81	12.18	23.41	16.02	54.44	40.91	46.72	83.99	59.77	69.84	50.06	29.23	36.91	15.06	21.39	17.68	15.36	17.30	16.27
CAE-Ensemble	40.18	41.99	41.07	15.88	26.68	19.91	55.98	41.52	47.68	88.61	59.90	71.48	51.67	29.92	37.90	18.41	19.51	18.94	20.35	18.27	19.26
PUAD	39.91	42.38	41.11	15.20	27.20	19.50	54.91	41.19	47.07	85.68	58.05	69.21	50.76	29.58	37.38	16.37	21.85	18.72	20.04	17.60	18.74
D3R	40.73	42.21	41.46	15.73	26.89	19.85	55.32	42.64	48.16	84.13	61.85	71.29	51.24	30.91	38.56	15.78	19.33	17.38	19.99	20.32	20.15
IGCL	42.31	46.97	<b>44.52</b>	17.48	28.02	<b>21.53</b>	59.20	44.94	<b>51.09</b>	84.98	63.54	<b>72.71</b>	52.10	32.77	<b>40.23</b>	17.39	25.90	<b>20.81</b>	21.97	22.81	<b>22.38</b>

**Table 4: Parameter sensitivity**

Dataset	SWaT			MSL		
$K$	P	R	F1	P	R	F1
4	84.67	61.93	71.54	20.01	20.17	20.09
8	84.23	62.48	71.74	20.38	21.42	20.89
12	84.52	62.97	72.17	20.50	21.51	20.99
16	84.21	63.19	72.20	20.43	21.69	21.04
20	84.09	63.25	72.19	20.33	21.65	20.97
24	84.98	63.54	<b>72.71</b>	21.97	22.81	<b>22.38</b>

window. They cannot explicitly learn the abnormal temporal dependencies from normal to anomaly precursors and thus have limited accuracy. The experiments demonstrate that learning the temporal dependencies from a pair of successive time series sub-sequences that change from the normal to the anomaly precursor is crucial to the anomaly prediction.

### 4.3 Parameter Sensitivity

We evaluate the impact of  $K$ , the size of the memory bank. The experimental results are shown in Table 4. If we set a bigger size for the memory bank to store more representative anomaly precursors, our model will have better results. When  $K$  increases to 24, our results are significantly better than the baselines, and  $K = 24$  is far

**Table 5: Larger look-forward windows**

Dataset	SWaT					
$f$	4		8		12	
Metric	F1	AUC	F1	AUC	F1	AUC
CAE-Ensemble	71.48	81.71	68.62	74.92	62.70	73.32
D3R	71.29	80.62	68.51	75.94	62.43	71.76
IGCL	<b>72.71</b>	<b>82.39</b>	<b>70.02</b>	<b>77.75</b>	<b>64.56</b>	<b>74.88</b>

smaller than  $O(2^n)$ , where  $n$ , i.e., the total number of variables, is 51 and 55 for SWaT and MSL datasets, respectively.

We evaluate the anomaly prediction with larger look-forward windows  $f$  in Table 5. We can see that IGCL still performs better than state-of-the-art baselines, even though the farther future anomalies are more difficult to predict.

The impact of the size of look-back windows  $h$  is in Table 6. We can see that IGCL is relatively stable to  $h$  and time series from different domains require different look-back windows to better identify the anomaly precursors.

### 4.4 Ablation Studies

We conduct ablation studies to validate the effectiveness of our key components. In particular, we compare IGCL with the following variants:

**Table 6: Parameter sensitivity**

Dataset	SWaT				PSM				SWAN			
$h$	P	R	F1	AUC-ROC	P	R	F1	AUC-ROC	P	R	F1	AUC-ROC
8	83.13	62.75	71.52	81.33	41.56	46.83	44.04	65.67	58.13	43.11	49.51	65.99
16	83.59	62.86	71.76	81.72	42.31	46.97	44.52	65.75	59.20	44.94	51.09	68.92
32	84.98	63.54	72.71	82.39	42.14	46.61	44.26	65.70	59.31	44.07	50.57	66.91
64	87.67	61.36	72.19	82.33	42.59	45.10	43.81	65.18	59.85	44.03	50.73	68.57

**Table 7: Ablation study.**

Dataset	PSM			SWAN			SWaT			GECCO		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1
w/o APPGM	41.58	40.51	41.04	55.82	41.03	47.30	84.19	60.28	70.26	51.88	29.03	37.23
w/o regularization loss	41.39	43.70	42.51	57.14	41.81	48.29	84.45	63.01	72.17	52.03	31.90	39.55
w/o memory bank	41.42	42.86	42.13	56.24	41.50	47.76	84.59	61.25	71.05	52.27	31.24	39.11
w Transformer	41.07	44.23	42.59	57.08	42.63	48.80	84.26	62.86	72.00	52.20	32.12	39.77
<b>IGCL</b>	42.31	46.97	<b>44.52</b>	59.20	44.94	<b>51.09</b>	84.98	63.54	<b>72.71</b>	52.10	32.77	<b>40.23</b>

- w/o anomaly precursor pattern generation module (APPGM): This variant directly uses random Gaussian noises as anomaly precursor patterns without the diffusion process.
- w/o regularization loss: This variant does not have regularization loss to control the diffusion process.
- w/o memory bank: This variant does not use the memory bank, and thus precursor patterns are generated during each iteration and appear in individual variable.
- w Transformer: This variant does not use the overlapping window-based TCN. It uses naive Transformer [41] to learn representations across timestamps.

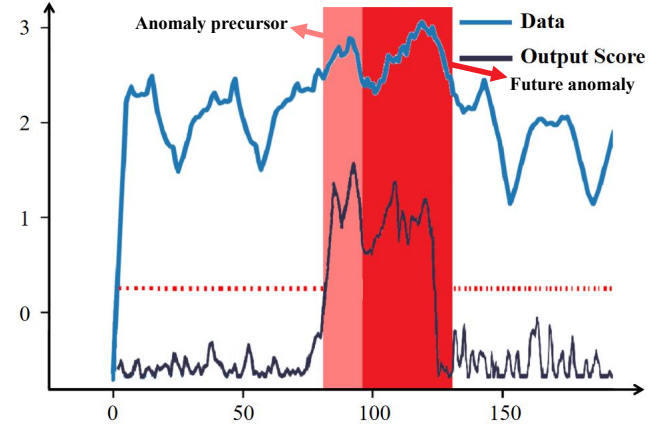
Table 7 shows the accuracy of different variants. From Table 7 we observe that: (1) As our anomaly precursor pattern generation module, regularization loss, and memory bank all help generate more complicated and representative anomaly precursors as negative samples, our model can have more accurate results. (2) Our overlapping window-based TCN can effectively learn the temporal dependencies from pairs of successive sub-sequences, and it is also more efficient than naive Transformer as shown in the complexity analysis.

#### 4.5 Case Study and Visualization

We visualize the anomaly prediction of our IGCL method on the SMAP dataset, as shown in Figure 4. By identifying the anomaly precursor that just begins to deviate from the normal, our model can report a warning at timestamp 80, ahead of the severe anomalies after timestamp 100.

### 5 Related Works

We review the relevant works on time series anomaly prediction, anomaly detection, contrastive learning, and generative learning. We also compare our method with related works in Table 8.



**Figure 4: IGCL can output larger scores on the precursors that start to deviate from the normal ahead of the severer anomalies.**

#### 5.1 Time Series Anomaly Prediction

Based on the current time series data, anomaly prediction aims to make the real-time judgment on whether there are starting anomaly precursors where data begin to deviate from normal to future anomalies. Anomaly prediction plays an important role in many scenarios that prioritize the safety or economic influence of anomalies. However, there are a few studies [2–4, 7, 17, 21, 32, 47] related to this problem. FEP [32] extracts statistical information from manually labeled frequency-based features as the starting anomaly precursors to predict future anomalies. EADS [2, 3] and VFP [17] use graphs to extract correlations among variables to help classify the starting anomaly precursors related to different variables. MCDA [7] proposes multi-instance attention, dCNN [4] proposes deep convolutional neural networks and CHE [47] proposes hierarchical windows, to extract temporal dependency features. They



**Table 8: Comparisons between representative methods that can be applied to unsupervised time series anomaly prediction.**

Methods	For time series	Unsupervised	Precursor Generation	Precursor Dependencies
[5, 22, 34]	✗	✓	✗	✗
[2, 3, 32]	✓	✗	✗	✗
[4, 7, 17, 47]	✓	✗	✗	✓
[6, 9, 12, 14, 15, 24, 31, 35–38] [21, 23, 30, 42, 46, 48, 50, 56]	✓	✓	✗	✗
Our IGCL	✓	✓	✓	✓

learn temporal dependencies from the pair of the normal time series sub-sequences and the starting anomaly precursors which help better identify starting abnormal changes for anomaly prediction.

However, these existing methods heavily rely on supervised learning of the temporal dependencies to get accurate anomaly prediction results. They are limited and fall short in the unsupervised time series anomaly prediction task for practical applications.

## 5.2 Time Series Anomaly Detection

There has been a lot of research on time series anomaly detection. Early works [55] utilize classification-based methods to detect anomalies in the historical time series data. Recent methods mainly focus on unsupervised settings based on the one-class classification theory [34]. They can be further categorized into four categories [51], including density estimation-based methods [5], clustering-based methods [51], autoregression and reconstruction-based methods [38] methods, and contrastive-based methods [50].

The clustering-based methods, such as K-means [35], DeepSVDD [34], and ITAD [37], aim to cluster normal samples within a center. THOC [36] uses a hierarchical RNN to learn dynamic clustering centers for time series data.

The density estimation-based anomaly detection methods, such as LOF [5], assume abnormal samples lie in a low-density region. IForest [22] uses the tree model and DAGMM [56] uses the encoding probability model to measure the density or distribution. LSTM [35] proposes forecasting-based approaches that forecast values and based on the forecasting errors detect whether the time points are anomalies. Autoencoder [14, 15, 24] (AE) is a classic architecture used in reconstruction. Omni [38], InterFusion [31] and LSTM-VAE [23] utilize the Variational AE and use reconstruction probability to detect anomalies. GANomaly [12] introduces a discriminator to generate normal data and detect anomalies based on reconstruction errors. CAE-Ensemble [6] proposes to ensemble different AEs for reconstruction. PUAD [30] uses meta-learning-oriented AE to reconstruct typical time series.

**Generative Learning.** The Generative Adversarial Network (GAN) is one kind of generative model, composed of a generator and a discriminator [27]. By designing the minimax theoretical game, the generator and discriminator improve each other and learn the underlying data distribution in an unsupervised setting. The diffusion-based network is a generative model that generates data with denoising steps [46]. ImDiffusion [9], DiffAD [46] and D3R [42] utilize diffusion-based networks to generate normal time series data and detect anomalies based on reconstruction errors. However, these methods only learn the underlying normal data distribution, but cannot generate anomaly precursors.

**Contrastive Learning.** Contrastive learning learns useful representations by contrasting positive pairs against negative pairs with data augmentation [8]. Recently, contrastive learning has been proposed for anomaly detection, such as DCdetector [50], where two view representations of a normal timestamp should be similar, and contrastive errors are used to detect anomalies. ATransformer [48] combines the reconstruction errors and contrastive errors together to detect anomalies.

PAD [21] proposes to utilize unsupervised anomaly detection methods and adjust specific thresholds on the anomaly scores to identify the anomaly precursors for anomaly prediction. However, these methods do not have negative samples and only reconstruct or contrast positive samples. They cannot learn the temporal precursor dependencies from the pair of normal time series and the anomaly precursor. Their performance is limited compared to supervised learning with labeled anomaly precursors as negative samples [4, 7, 17, 47].

## 6 CONCLUSION

We present an importance-based generative contrastive learning model for a novel problem of unsupervised time series anomaly prediction. We propose anomaly precursor pattern generation, with diffusion-based transformation and variance regularization, to generate diverse anomaly precursors as labeled negative data. Then, we propose an overlapping window-based contrastive learning to distinguish anomaly precursors from the normal efficiently. Last, we propose a memory bank with importance scores to store representative anomaly precursors to generate complicated precursors efficiently. Experiments on seven benchmark datasets demonstrate the superiority of our method. In future works, it is worth improving IGCL with pre-training on large-scale time series data.

## Acknowledgments

This work was partially supported by National Natural Science Foundation of China (62372179), Independent Research Fund Denmark (8022-00246B), and the VILLUM FONDEN (40567).

## References

- [1] Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. 2021. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *SIGKDD*. 2485–2494.
- [2] Yihao Ang, Qiang Huang, Anthony Tung, and Zhiyong Huang. 2023. A Stitch in Time Saves Nine: Enabling Early Anomaly Detection with Correlation Analysis. In *ICDE*. IEEE, 1832–145.
- [3] Yihao Ang, Qiang Huang, Anthony KH Tung, and Zhiyong Huang. 2024. EADS: An Early Anomaly Detection System for Sensor-Based Multivariate Time Series. In *ICDE*. IEEE, 5433–5436.
- [4] Paul Boniol, Mohammed Meftah, Emmanuel Remy, Bruno Didier, and Themis Palpanas. 2023. dCNN/dCAM: anomaly precursors discovery in multivariate

- time series with deep convolutional neural networks. *Data-Centric Engineering* 4 (2023).
- [5] Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. In *SIGMOD*.
  - [6] David Campos, Tung Kieu, Chenjuan Guo, Feiteng Huang, Kai Zheng, Bin Yang, and Christian S. Jensen. 2022. Unsupervised Time Series Outlier Detection with Diversity-Driven Convolutional Ensembles. *PVLDB* 15, 3 (2022).
  - [7] Vitor Cerqueira, Lu's Torgo, and Carlos Soares. 2023. Early anomaly detection in time series: a hierarchical approach for predicting critical health episodes. *Mach. Learn.* 112, 11 (2023), 4409–4430.
  - [8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *ICML*, Vol. 119.
  - [9] Yuhang Chen, Chaoyun Zhang, Minghua Ma, Yudong Liu, Ruomeng Ding, Bowen Li, Shilin He, Saravan Rajmohan, Qingwei Lin, and Dongmei Zhang. 2023. ImDiffusion: Imputed Diffusion Models for Multivariate Time Series Anomaly Detection. In *Proc. VLDB*.
  - [10] Yunyao Cheng, Peng Chen, Chenjuan Guo, Kai Zhao, Qingsong Wen, Bin Yang, and Christian S. Jensen. 2024. Weakly Guided Adaptation for Robust Time Series Forecasting. *Proc. VLDB Endow.* (2024).
  - [11] Razvan-Gabriel Cirstea, Bin Yang, and Chenjuan Guo. 2019. Graph Attention Recurrent Neural Networks for Correlated Time Series Forecasting. In *KDD*.
  - [12] Bowen Du, Xuanxuan Sun, Junchen Ye, Ke Cheng, Jingyuan Wang, and Leilei Sun. 2023. GAN-Based Anomaly Detection for Multivariate Time Series Using Polluted Training Set. *IEEE Trans. Knowl. Data Eng.* 35, 12 (2023).
  - [13] Nicolai Engelmann and Heinz Koepl. 2022. Forward-Backward Latent State Inference for Hidden Continuous-Time semi-Markov Chains. In *NeurIPS*.
  - [14] Yuchen Fang, Jiandong Xie, Yan Zhao, Lu Chen, Yunjun Gao, and Kai Zheng. 2024. Temporal-Frequency Masked Autoencoders for Time Series Anomaly Detection. In *ICDE*.
  - [15] Xinli Hao, Yile Chen, Chen Yang, Zhihui Du, Chaohong Ma, Chao Wu, and Xiaofeng Meng. 2024. From Chaos to Clarity: Time Series Anomaly Detection in Astronomical Observations. In *ICDE*.
  - [16] Xiao He, Ye Li, Jian Tan, Bin Wu, and Feifei Li. 2023. OneShotSTL: One-Shot Seasonal-Trend Decomposition for Online Time Series Anomaly Detection And Forecasting. In *Proc. VLDB*.
  - [17] Hadi Hojjati, Mohammadreza Sadeghi, and Narges Armanfard. 2023. Multivariate Time-Series Anomaly Detection with Temporal Self-supervision and Graphs: Application to Vehicle Failure Prediction. In *PKDD*, Vol. 14175.
  - [18] Alexis Huet, Jos'e Manuel Navarro, and Dario Rossi. 2022. Local Evaluation of Time Series Anomaly Detection Algorithms. In *SIGKDD*.
  - [19] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Soderstrom. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *SIGKDD*. 387–395.
  - [20] Arun Shankar Iyer, J. Saketha Nath, and Sunita Sarawagi. 2014. Maximum Mean Discrepancy for Class Ratio Estimation: Convergence Bounds and Kernel Selection. In *ICML*, Vol. 32. JMLR.org, 530–538.
  - [21] Sheo Yon Jhin, Jaehoon Lee, and Noseong Park. 2023. Precursor-of-Anomaly Detection for Irregular Time Series. In *SIGKDD*. ACM, 917–929.
  - [22] Pawel Karczmarek, Adam Kiersztyn, Witold Pedrycz, and Ebru Al. 2020. K-Means-based isolation forest. *Knowledge-based systems* 195 (2020), 105659.
  - [23] Tung Kieu, Bin Yang, Chenjuan Guo, Razvan-Gabriel Cirstea, Yan Zhao, Yale Song, and Christian S. Jensen. 2022. Anomaly Detection in Time Series with Robust Variational Quasi-Recurrent Autoencoders. In *ICDE*. 1342–1354.
  - [24] Tung Kieu, Bin Yang, Chenjuan Guo, Christian S. Jensen, Yan Zhao, Feiteng Huang, and Kai Zheng. 2022. Robust and Explainable Autoencoders for Unsupervised Time Series Outlier Detection. In *ICDE*. 3038–3050.
  - [25] Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. 2021. Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift. In *ICLR*. OpenReview.net.
  - [26] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*, Yoshua Bengio and Yann LeCun (Eds.).
  - [27] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *ICLR*.
  - [28] Lukas Kohs, Bastian Alt, and Heinz Koepl. 2022. Markov Chain Monte Carlo for Continuous-Time Switching Dynamical Systems. In *ICML*.
  - [29] Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. 2021. Revisiting Time Series Outlier Detection: Definitions and Benchmarks. In *NeurIPS*.
  - [30] Yuxin Li, Wenchao Chen, Bo Chen, Dongsheng Wang, Long Tian, and Mingyuan Zhou. 2023. Prototype-oriented unsupervised anomaly detection for multivariate time series. In *ICML*, Vol. 202. PMLR, 19407–19424.
  - [31] Zhihan Li, Youjian Zhao, Jiaqi Han, Ya Su, Rui Jiao, Xidao Wen, and Dan Pei. 2021. Multivariate Time Series Anomaly Detection and Interpretation using Hierarchical Inter-Metric and Temporal Embedding. In *SIGKDD*.
  - [32] Aleksandra Nina. 2024. Analysis of VLF Signal Noise Changes in the Time Domain and Excitations/Attenuations of Short-Period Waves in the Frequency Domain as Potential Earthquake Precursors. *Remote. Sens.* 16, 2 (2024).
  - [33] John Paparrizos, Paul Boniol, Themis Palpanas, Ruey Tsay, Aaron J. Elmore, and Michael J. Franklin. 2022. Volume Under the Surface: A New Accuracy Evaluation Measure for Time-Series Anomaly Detection. *Proc. VLDB* (2022).
  - [34] Lukas Ruff, Robert A. Vandermeulen, Nico G'ornitz, Lucas Deecke, Shoaib A. Siddiqui, Alexander Binder, Emmanuel M'uller, and Marius Kloft. 2018. Deep One-Class Classification. In *ICML*, Vol. 80. 4393–4402.
  - [35] Sebastian Schmidl, Felix Naumann, and Thorsten Papenbrock. 2024. AutoTSAD: Unsupervised Holistic Anomaly Detection for Time Series Data. In *Proc. VLDB*.
  - [36] Lifeng Shen, Zhuocong Li, and James Kwok. 2020. Timeseries anomaly detection using temporal hierarchical one-class network. *NeurIPS* (2020).
  - [37] Youjin Shin, Sangyup Lee, Shahroz Tariq, Myeong Shin Lee, Okchul Jung, Daewon Chung, and Simon S. Woo. 2020. ITAD: Integrative Tensor-based Anomaly Detection System for Reducing False Positives of Satellite Systems. In *CIKM*.
  - [38] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *SIGKDD*. 2828–2837.
  - [39] Yuting Sun, Guansong Pang, Guanhua Ye, Tong Chen, Xia Hu, and Hongzhi Yin. 2024. Unraveling the 'Anomaly' in Time Series Anomaly Detection: A Self-supervised Tri-domain Solution. In *ICDE*.
  - [40] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2016. Instance Normalization: The Missing Ingredient for Fast Stylization. *arXiv/1607.08022* (2016). arXiv:1607.08022
  - [41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*. 5998–6008.
  - [42] Chengsen Wang, Zirui Zhuang, Qi Qi, Jingyu Wang, Xingyu Wang, Haifeng Sun, and Jianxin Liao. 2023. Drift doesn't Matter: Dynamic Decomposition with Diffusion Reconstruction for Unstable Multivariate Time Series Anomaly Detection. In *NeurIPS*.
  - [43] Rui Wang, Xudong Mou, Renyu Yang, Kai Gao, Pin Liu, Chongwei Liu, Tianyu Wo, and Xudong Liu. 2024. CutAddPaste: Time Series Anomaly Detection by Exploiting Abnormal Knowledge. In *SIGKDD*.
  - [44] Renjie Wu and Eamonn J. Keogh. 2022. Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress. In *ICDE*.
  - [45] Xinle Wu, Dalin Zhang, Miao Zhang, Chenjuan Guo, Bin Yang, and Christian S. Jensen. 2023. AutoCTS+: Joint Neural Architecture and Hyperparameter Search for Correlated Time Series Forecasting. *Proc. ACM Manag. Data* 1, 1 (2023).
  - [46] Chunjing Xiao, Zehua Gou, Wenxin Tai, Kunpeng Zhang, and Fan Zhou. 2023. Imputation-based Time-Series Anomaly Detection with Conditional Weight-Incremental Diffusion Models. In *SIGKDD*. 2742–2751.
  - [47] Dongkuan Xu, Wei Cheng, Jingchao Ni, Dongsheng Luo, Masanao Natsumeda, Dongjin Song, Bo Zong, Haifeng Chen, and Xiang Zhang. 2021. Deep Multi-Instance Contrastive Learning with Dual Attention for Anomaly Precursor Detection. In *SIAM SDM*.
  - [48] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. 2022. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. In *ICLR*.
  - [49] Ronghui Xu, Hao Miao, Senzhang Wang, Philip S Yu, and Jianxin Wang. 2024. PeFAD: a parameter-efficient federated framework for time series anomaly detection. In *SIGKDD*. 3621–3632.
  - [50] Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. 2023. DCdetector: Dual Attention Contrastive Representation Learning for Time Series Anomaly Detection. In *SIGKDD*. ACM.
  - [51] Aoqian Zhang, Shuqing Deng, Dongping Cui, Ye Yuan, and Guoren Wang. 2023. An Experimental Evaluation of Anomaly Detection in Time Series. In *Proc. VLDB*.
  - [52] Kai Zhao, Ting Bai, Bin Wu, Bai Wang, Youjie Zhang, Yuanyu Yang, and Jian-Yun Nie. 2020. Deep adversarial completion for sparse heterogeneous information network embedding. In *WWW*. 508–518.
  - [53] Kai Zhao, Chenjuan Guo, Yunyao Cheng, Peng Han, Miao Zhang, and Bin Yang. 2023. Multiple Time Series Forecasting with Dynamic Graph Modeling. *Proc. VLDB Endow.* 17, 4 (2023), 753–765.
  - [54] Kai Zhao, Yukun Zheng, Tao Zhuang, Xiang Li, and Xiaoyi Zeng. 2022. Joint Learning of E-commerce Search and Recommendation with a Unified Graph Neural Network. In *WSDM*. 1461–1469.
  - [55] Yan Zhao, Liwei Deng, Xuanhao Chen, Chenjuan Guo, Bin Yang, Tung Kieu, Feiteng Huang, Torben Bach Pedersen, Kai Zheng, and Christian S. Jensen. 2022. A Comparative Study on Unsupervised Anomaly Detection for Time Series: Experiments and Analysis. *CoRR* abs/2209.04635 (2022).
  - [56] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Dae-ki Cho, and Haifeng Chen. 2018. Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. In *ICLR*. OpenReview.net.

## A Proof

**Theorem 1:** Under the unsupervised setting, the future time series anomaly is unpredictable if we cannot identify the anomaly precursor in the current observation that follows the same distribution as the normal previous sub-sequences do.

**Proof 1:** As the current observations follow the same conditional distribution as the normal sub-sequence follows, we have  $\mathcal{N} = \mathcal{A}$ . For any model  $\mathcal{F}$  we have:

$$\mathbb{P}_{\mathbf{H}_t \sim \mathcal{A}} [\mathcal{F}(\mathbb{X}_{t,h}^b) = \hat{p}_t] - \mathbb{P}_{\mathbf{H}_t \sim \mathcal{N}} [\mathcal{F}(\mathbb{X}_{t,h}^b) = \hat{p}_t] = 0. \quad (16)$$

Thus, the predicted probability score for the anomaly precursor in the current observation is equal to that for the normal sub-sequences. Therefore, no unsupervised model can predict the future anomaly if we cannot identify the anomaly precursors in the current observations when the anomaly precursor follows the same distribution as the normal sub-sequences do.  $\square$

**Theorem 2:** Under the unsupervised setting, the future time series anomaly is predictable if we can identify its anomaly precursor in the current observation that follow a distribution different from the normal sub-sequences.

**Proof 2:** As the anomaly precursor follows a data distribution different from the normal sub-sequences, we have  $\mathcal{N} \neq \mathcal{A}$ . According to the Maximum Mean Discrepancy (MMD) theory [20], there exists at least one mapping function  $\mathcal{F}$  satisfying:

$$\mathbb{E}_{\mathbf{H}_t \sim \mathcal{A}} [\mathcal{F}(\mathbb{X}_{t,h}^b)] - \mathbb{E}_{\mathbf{H}_t \sim \mathcal{N}} [\mathcal{F}(\mathbb{X}_{t,h}^b)] > 0 \quad (17)$$

Thus, the expectation of the predicted probability score for the anomaly precursor is larger than that for the normal sub-sequence. Therefore, there exists at least one model  $\mathcal{F}$  that can predict a larger probability score to identify that there is the starting anomaly precursor and the future time series is more like an anomaly when the anomaly precursor follows the data distribution different from the normal sub-sequences.  $\square$

Based on the above analysis, we can only predict the future time series anomalies whose anomaly precursors just begin to deviate from the nearest previous normal sub-sequences under the unsupervised setting. However, we do not have labeled anomaly precursor data for the model to learn the temporal dependencies to estimate the Markov conditional distributions for  $\mathcal{A} = P(\cdot | \mathbf{H}_{t-h-1}, \mathbf{H}_t \text{ is anomaly precursor})$  and  $\mathcal{N} = P(\cdot | \mathbf{H}_{t-h-1}, \mathbf{H}_t \text{ is normal})$ . Thus, the best potential model is that we generate different kinds of potential anomaly precursors to simulate the beginning of deviation from normal and then learn the temporal dependency features.

## B The pollution process

The noise pollution process  $q(x^s | x^{s-1})$  is shown as

$$x^s = \sqrt{1 - \beta^s} x^{s-1} + \sqrt{\beta^s} \epsilon^s, \quad \forall s \in \{1, 2, \dots, S\}, \quad (18)$$

where  $\epsilon^s \sim N(0, I)$  are the independent Gaussian pollutions to be added in different steps,  $I$  is the identity matrix,  $0 < \beta^s < 1$  are the hyper-parameters control the amount of pollution added in each step. Thus, we have:

$$x^s = \sqrt{1 - \beta^s} x^{s-1} + \sqrt{\beta^s} \epsilon^s \quad (19)$$

$$\begin{aligned} &= \sqrt{1 - \beta^s} (\sqrt{1 - \beta^{s-1}} x^{s-2} + \sqrt{\beta^{s-1}} \epsilon^{s-1}) + \sqrt{\beta^s} \epsilon^s \\ &= \dots \\ &= (\sqrt{1 - \beta^s} \sqrt{1 - \beta^{s-1}} \dots \sqrt{1 - \beta^1}) x^0 + \\ &\quad (\sqrt{1 - \beta^s} \sqrt{1 - \beta^{s-1}} \dots \sqrt{1 - \beta^2}) \sqrt{\beta^1} \epsilon^1 + \\ &\quad (\sqrt{1 - \beta^s} \sqrt{1 - \beta^{s-1}} \dots \sqrt{1 - \beta^3}) \sqrt{\beta^2} \epsilon^2 + \dots + \\ &\quad \sqrt{1 - \beta^s} \sqrt{\beta^{s-1}} \epsilon^{s-1} + \sqrt{\beta^s} \epsilon^s \end{aligned}$$

For any independent Gaussian distributions  $\epsilon_1 \sim N(\mu_1, \sigma_1^2)$  and  $\epsilon_2 \sim N(\mu_2, \sigma_2^2)$ , we have the following properties:

$$a\epsilon_1 + b \sim N(a\mu_1 + b, a^2\sigma_1^2), \quad (20)$$

$$\epsilon_1 + \epsilon_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2). \quad (21)$$

We also have:

$$\begin{aligned} 1 &= (\sqrt{1 - \beta^s} \sqrt{1 - \beta^{s-1}} \dots \sqrt{1 - \beta^1})^2 + \\ &\quad (\sqrt{1 - \beta^s} \sqrt{1 - \beta^{s-1}} \dots \sqrt{1 - \beta^2})^2 \beta^1 + \\ &\quad (\sqrt{1 - \beta^s} \sqrt{1 - \beta^{s-1}} \dots \sqrt{1 - \beta^3})^2 \beta^2 + \dots + \\ &\quad (1 - \beta^s) \beta^{s-1} + \beta^s, \end{aligned} \quad (22)$$

Thus, we have:

$$x^s = \sqrt{\alpha^s} x^0 + \sqrt{1 - \alpha^s} \epsilon, \quad (23)$$

where  $\epsilon \sim N(0, I)$  and  $\sqrt{\alpha^s} = \sqrt{1 - \beta^s} \sqrt{1 - \beta^{s-1}} \dots \sqrt{1 - \beta^1}$ . Then, for enough  $S$  steps and proper  $0 < \beta^s < 1$ ,  $\sqrt{\alpha^s}$  will be close to 0, and the output from the noise pollution process, i.e.,  $x^S$ , will be gradually corrupted into random Gaussian distribution.

## C Regularization loss

$$\begin{aligned} &KL(N(\mu, \sigma^2) \| N(\mu, I)) \\ &= \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \left( \log \frac{e^{-(x-\mu)^2/2\sigma^2} / \sqrt{2\pi\sigma^2}}{e^{-(x-\mu)^2/2} / \sqrt{2\pi}} \right) dx \\ &= \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \log \left\{ \frac{1}{\sqrt{\sigma^2}} \exp \left\{ \frac{1}{2} (x-\mu)^2 \left( 1 - \frac{1}{\sigma^2} \right) \right\} \right\} dx \\ &= \frac{1}{2} \int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2} \left[ -\log \sigma^2 + (x-\mu)^2 \left( 1 - \frac{1}{\sigma^2} \right) \right] dx \\ &= \frac{1}{2} \left( -\log \sigma^2 + \sigma^2 \left( 1 - \frac{1}{\sigma^2} \right) \right) \\ &= \frac{1}{2} \left( -\log \sigma^2 + \sigma^2 - 1 \right). \end{aligned} \quad (24)$$

## D Experimental details

### D.1 Datasets

- SMD [38] (Server Machine Dataset) is collected from a large computing cluster, where different server machines stack accessed traces of resource utilization with 38 dimensions.
- PSM [1] (Pooled Server Metrics) is a 25-dimension dataset collected by eBay that shows the performance of multiple application servers.
- MSL [19] (Mars Science Laboratory Rover) is from the spacecraft monitoring systems and collected by NASA. It shows the health check-up data of the sensors from the Mars rover.
- SMAP [19] (Soil Moisture Active Passive) is a 25-dimension dataset collected by NASA, which contains soil samples and telemetry information.

**Table 9: More experiment results of state-of-the-art methods, presented in percentages.**

Dataset	PSM	SMAP	SWAN	SWaT	GECCO	SMD	MSL	UCR				Credit			
Metrics	AUC-ROC							P	R	F1	AUC-ROC	P	R	F1	AUC-ROC
DCdetector	50.91	43.98	51.00	65.57	50.79	50.44	30.34	12.59	10.90	11.68	48.65	37.50	12.96	19.26	64.34
PAD	51.02	50.13	50.85	61.35	49.71	49.66	48.28	29.92	20.35	24.22	51.67	31.49	16.01	21.23	68.76
ATransformer	54.91	53.70	51.95	75.63	50.39	59.30	50.21	32.07	20.86	25.28	51.68	17.58	26.85	21.25	68.55
CAE-Ensemble	63.68	58.76	65.41	81.71	52.94	67.99	51.26	34.15	21.47	26.37	58.26	18.16	45.98	26.04	82.16
D3R	63.93	59.61	66.13	80.62	53.09	66.83	51.83	39.10	20.77	27.13	60.11	17.85	50.17	26.33	80.68
<b>IGCL</b>	<b>65.75</b>	<b>61.92</b>	<b>68.92</b>	<b>82.39</b>	<b>54.02</b>	<b>69.97</b>	<b>55.89</b>	36.43	31.04	<b>33.52</b>	<b>69.82</b>	17.62	59.07	<b>27.15</b>	<b>84.00</b>

**Table 10: Anomaly prediction for irregular time series**

Dataset	SWaT											
Dropping ratio	0%				30%				50%			
Metric	P	R	F1	AUC	P	R	F1	AUC	P	R	F1	AUC
PAD	74.83	59.09	66.04	61.35	74.17	57.80	64.97	60.71	69.29	50.94	58.71	56.09
IGCL+	84.98	63.54	<b>72.71</b>	<b>82.39</b>	84.02	59.97	<b>69.98</b>	<b>78.63</b>	70.81	53.77	<b>61.13</b>	<b>60.17</b>

- SWAN [29] is extracted from solar photospheric vector magnetograms in the Spaceweather, which is the benchmark dataset used in NeurIPS Competition Track.
- SWaT [48] (Secure Water Treatment) is collected from sensors of the critical infrastructure systems under continuous operations for secure water treatment.
- GECCO [29] is a 9-dimension dataset collected from the cyber-physical systems showing the drinking water quality.
- UCR [44] is a 1-dimension dataset containing 250 sub-datasets from various domains.
- Credit [55] is from the finance scenario where the anomaly rate is only 0.172%.

These datasets are representative open benchmarks for multi-variable and uni-variable time series anomaly tasks, and thus we evaluate the unsupervised time series anomaly prediction problem on these datasets. More statistical information and details can be seen in Table 2, where  $N$  is the number of variables observed at each timestamp. We follow the same training-validation-test splits as in the original papers [1, 6, 19, 29, 38, 48] shown in each column.

## D.2 Settings

We employ the official open-source implementations of baseline methods and have carefully tuned their hyper-parameters based on the recommendations from the original papers. The target label from the look-forward window is modified for the anomaly prediction task following PAD [21]. The anomaly score from the current look-back window is used for identifying the anomaly precursors to predict future anomalies also following PAD [21]. The experiments

are conducted on an Ubuntu 18.04.5 LTS system server, with Intel(R) Xeon(R) Gold 5215 CPU @ 2.50GHz and NVIDIA Quadro RTX 8000 GPU. All the deep learning-based models are executed with Pytorch 1.2.0. The size of the default look-forward window size  $f$  is 4, and we also evaluate with larger look-forward windows of {4, 8, 12}. We use Adam optimizer with the default parameter and the learning rate is default 0.0001. The batch size is tuned from {32, 64}. The look-back window size  $h$  and the number of sub-sequences  $b$  are tuned from {8, 16, 32, 64}. The memory bank size  $K$  is tuned from {16, 20, 24} and we evaluate with more sizes in the parameter study. The number of positive samples  $P$  is tuned from {12, 16, 20}. The  $\lambda$  which controls the strength of regularization is tuned from {0.5, 1}. We use three parallel kernel sizes {2, 3, 5} for our overlapping TCN. Following existing works [38, 55], we choose the best threshold for all methods.

## D.3 More experiment results

Table 9 shows more experiment results of most state-of-the-art methods on all datasets using more metrics.

## D.4 Irregular time series

PAD uses interpolation and neural controlled differential equations (NCDE) to support irregular time series. The interpolation method can be applied before any models. Our TCN backbone can also be replaced with NCDE. We show the experimental results on the irregular series by dropping random points following PAD. We only add the interpolation method before our model, which is denoted as IGCL+. We can see that our method can be extended to irregular series easily and still perform well.