

SimAD: A Simple Dissimilarity-based Approach for Time Series Anomaly Detection

Zhijie Zhong, Zhiwen Yu*, Senior Member IEEE, Xing Xi, Yue Xu, Wenming Cao, Member IEEE, Yiyuan Yang, Kaixiang Yang, Member IEEE, Jane You

Abstract—Despite the prevalence of reconstruction-based deep learning methods, time series anomaly detection remains a tremendous challenge. Existing approaches often struggle with limited temporal contexts, insufficient representation of normal patterns, and flawed evaluation metrics, all of which hinder their effectiveness in detecting anomalous behavior. To address these issues, we introduce a Simple dissimilarity-based approach for time series Anomaly Detection, referred to as **SimAD**. Specifically, SimAD first incorporates a patching-based feature extractor capable of processing extended temporal windows and employs the EmbedPatch encoder to fully integrate normal behavioral patterns. Second, we design an innovative ContrastFusion module in SimAD, which strengthens the robustness of anomaly detection by highlighting the distributional differences between normal and abnormal data. Third, we introduce two robust enhanced evaluation metrics, Unbiased Affiliation (UAff) and Normalized Affiliation (NAff), designed to overcome the limitations of existing metrics by providing better distinctiveness and semantic clarity. The reliability of these two metrics has been demonstrated by both theoretical and experimental analyses. Experiments conducted on seven diverse time series datasets clearly demonstrate SimAD’s superior performance compared to state-of-the-art methods, achieving relative improvements of 19.85% on F1, 4.44% on Aff-F1, 77.79% on NAff-F1, and 9.69% on AUC on six multivariate datasets. Code and pre-trained models are available at <https://github.com/EmorZz1G/SimAD>.

Index Terms—Data mining, Time series, Anomaly detection, Deep learning, Outlier detection, Evaluation metrics.

I. INTRODUCTION

Time series anomaly detection (TSAD) is a critical component of time series analysis, focused on accurately detecting abnormal patterns in time series data and identifying their specific locations [1, 2]. TSAD methods utilize time series data to identify anomalies in web traffic, which play a vital role in ensuring the stability, security, and efficient functioning of web

Zhijie Zhong is with Pengcheng Laboratory, Shenzhen, Guangdong, 518066, China, and also with the School of Future Technology, South China University of Technology, Guangzhou, Guangdong 510650, China.

Zhiwen Yu is with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong 510650, China, and also with the Pengcheng Laboratory, Shenzhen, Guangdong 518066, China. Email: zhwyu@scut.edu.cn. Telephone number: 86-20-62893506. Fax number: 86-20-39380288.

Xing Xi, Yue Xu and Kaixiang Yang are with the School of Computer Science and Engineering, South China University of Technology, Guangzhou, Guangdong 510650, China.

Wenming Cao is with the Department of Information and Computing Science, Chongqing Jiaotong University.

Yiyuan Yang, from the Department of Computer Science at the University of Oxford, Oxford, UK, worked down as an intern at Alibaba Group.

Jane You is with the Department of Industrial and Systems Engineering in The Hong Kong Polytechnic University, China. (Email: jane.you@polyu.edu.hk)

*Corresponding author: Zhiwen Yu.

services [3]. Unsupervised methods have garnered considerable attention in academic research, particularly for addressing this challenge through reconstruction-based approaches [4, 5, 6, 7, 8]. These methods assume that models are perfectly trained on normal data and assign higher anomaly scores to anomalous data during the testing phase. However, these methods have shown insufficient performance in practical applications. A thorough review of existing research [9, 10, 11, 12], coupled with comprehensive experiments (detailed in our analyses), has enabled us to identify several critical challenges in the field of time series anomaly detection:

- **Challenge 1:** Many methods rely on the reconstruction assumption, which is inadequate for enhancing the detection performance and may not always hold true [13, 9]. Our experiments, including both sensitivity and ablation analyses, validate this limitation.
- **Challenge 2:** Failure to adequately utilize extended time windows: The complexity of the attention mechanism has constrained previous methods, capping the window length at 200 or fewer [13, 14]. This limitation, in turn, prevents the capture of more informative data.
- **Challenge 3:** Limited expressive power hinders the representation of normal features. On one hand, certain methods fail to effectively model either normal or abnormal data, or both. On the other hand, most models are constrained by a limited number of parameters, which restricts their expressive capacity.

To address the above challenges, we propose a **Simple dissimilarity-based approach for time series Anomaly Detection** method, referred to as **SimAD**, in both univariate and multivariate settings. Specifically, to tackle **Challenge 2**, we design a feature extractor that can process longer time windows by splitting the sequence into multiple patches. This strategy enables SimAD to learn extended temporal receptive fields while using fewer parameters. To address **Challenges 1** and **3**, we design the EmbedPatch encoder and incorporate an enhanced attention mechanism for layer-wise modeling of dissimilarities between normal and abnormal data features. Thus, the EmbedPatch encoder can learn discriminative representations of normal data more effectively. Essentially, the proposed EmbedPatch encoder is a prototype, using V vectors to store prototypical features of normal data. The key differences between the EmbedPatch encoder and traditional prototypes can be highlighted in two aspects: 1) Prototypes typically use the last layer to preserve features, while the EmbedPatch encoder, acting as the value matrices for the attention mechanism, not only preserves features at different layers but also serves as a querying

repository for each level; 2) In traditional prototypes, query results are obtained by calculating the similarity between prototypes and features. In contrast, the EmbedPatch encoder generates query results by interacting the attention scores with the value matrices. With the EmbedPatch encoder, each layer of attention mechanism incorporates a series of embedded patches to learn features of that specific layer, giving it a stronger capability to learn richer and more distinctive representations, which are beneficial to performance improvement of anomaly detection. Finally, we introduce the ContrastFusion module to amplify the divergence of normal and abnormal data's distributions for further accentuating their dissimilarity.

Furthermore, we have analyzed the shortcomings of existing evaluation metrics, including inflated metrics, low discriminative power, and insufficient semantic relevance. These issues have contributed to false perceptions of progress in TSAD [15, 16, 17]. To enable a fair comparison of performance, we propose two improved evaluation metrics: Unbiased Affiliation (**UAff**) and Normalized Affiliation (**NAff**). These metrics are designed to address the limitations of existing metrics by providing better distinctiveness and semantics. We have conducted analyses from both theoretical and experimental perspectives to establish the reliability of the newly proposed metrics. As demonstrated via extensive experiments, the proposed approach outperforms the state-of-the-art (SOTA) methods in time series anomaly detection.

In summary, our paper makes the following contributions:

- 1) We propose **SimAD**, a simple yet effective algorithm designed for time series anomaly detection that can handle extended time windows. SimAD is a straightforward framework based on dissimilarity measures, and its effectiveness has been validated through comprehensive evaluations from multiple perspectives.
- 2) We introduce two improved TSAD evaluation metrics, **UAff** and **NAff**, which address challenges such as inaccurate assessments and lack of semantic clarity. The reliability of the proposed metrics is demonstrated through theoretical and experimental analyses. Furthermore, a comprehensive evaluation of the issues, limitations, and strengths of existing metrics is conducted.
- 3) Our algorithm outperforms significantly SOTA methods on real-world datasets, excelling in point-level evaluations such as F1 and AUC, as well as sequence-level evaluations like Aff, UAff and NAff (See Fig. 3). Moreover, we have released both our code and pre-trained models.

II. RELATED WORK

Unsupervised learning methods [4, 18] have garnered considerable attention due to the scarcity of labeled data for anomaly detection in time series sequences [1, 19, 13, 20, 9, 21]. In past studies, anomaly detection algorithms have become one of the key data analysis tools and are widely used in fields like remote sensing and imaging [22, 23, 24, 25]. These methods can be categorized as follows (detailed discussion in Appendix A): (1) Algorithms based on classical machine learning [26, 27, 28] transform traditional machine learning approaches into deep networks, enhancing their ability to handle complex data. (2)

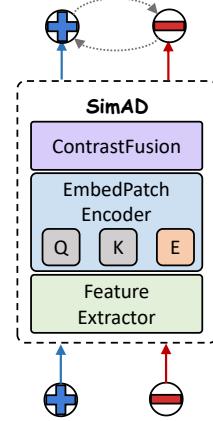


Fig. 1. The Overview of SimAD.

Reconstruction-based approaches [19, 14, 29] involve training models using normal data and leveraging reconstruction error as an anomaly score, attributing higher scores to anomalous data during testing. (3) Prediction-based techniques, as demonstrated in [30], learn from historical data to predict future observations, considering prediction errors as the foundation for anomaly detection. (4) Generative adversarial learning based methods [31, 32, 33] utilize generative models to learn the distribution of normal data and a discriminator network to detect anomalies.

Recently, some innovative algorithms have emerged in the field of anomaly detection, including: (1) Transformer-based approaches [13, 34, 21] leverage the power of Transformer, which has shown exceptional success in natural language processing tasks, and are increasingly being applied to anomaly detection. (2) Contrastive learning-based methods [35, 36, 20, 37, 38] utilize contrastive learning to obtain robust representations, specifically tailored for anomaly detection. (3) Diffusion-based methods [39, 40] model the propagation of anomalies in complex networks and time series sequences by using the diffusion process. (4) Large Language Models (LLMs) [41] exploit cutting-edge models, like GPT-2, adapted specifically for anomaly detection tasks, capitalizing on the sophisticated architectures and knowledge representation capabilities.

These algorithms can be considered extensions or variations of the above categories, which incorporate advanced network structures and knowledge representation methods to enhance anomaly detection performance. However, existing methods have not effectively addressed the above three challenges. Most methods are constrained by a limited number of learnable parameters, which hinders them from capturing long-term dependencies in data. Moreover, some reconstruction-based methods solely focus on the task of reconstruction. The differences between our dissimilarity-based approach and Contrastive learning-based methods are in the framework (EmbedPatch & ContrastFusion), motivation (dissimilarity of perspectives [42]), and implementation, as discussed in Appendix I.

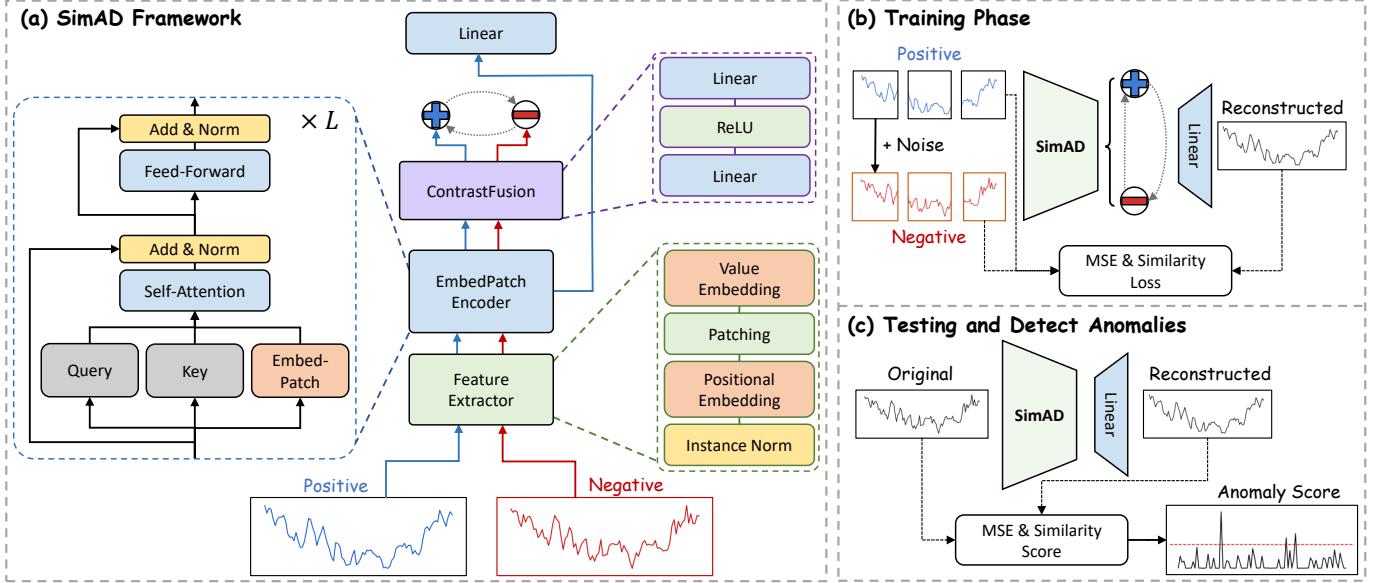


Fig. 2. The framework and workflow of SimAD.

III. PROPOSED APPROACH

A. Framework of SimAD

1) *Overview*: In Fig. 2(a), SimAD comprises a Feature Extractor, EmbedPatch Encoder, and ContrastFusion module. In the context of temporal anomaly detection, the original time series data is denoted as $\mathbf{X} \in \mathbb{R}^{T \times C}$, where T denotes the length of time series and C the number of channels. The goal of this task is to predict the label $y \in \mathbb{R}^T$ for \mathbf{X} , where $y_i = 1$ indicates an anomaly at the i -th time point and $y_i = 0$ otherwise. The input data \mathbf{X} is first processed by the Feature Extractor to generate patch tokens \mathbf{N} . Next, at each layer of the EmbedPatch Encoder, the attention mechanism combines the patch tokens from the previous layer with the current tokens to capture their interdependencies. We then utilize a linear layer to reconstruct the original features and compute anomaly scores. Finally, we design the ContrastFusion module to strengthen the distinction between normal and abnormal data distributions, thereby enhancing overall detection performance.

2) *Feature Extractor*: To extract unified sequence-level temporal features, we devised a patch-based feature extractor, which allows SimAD to process longer time windows and capture richer semantic information. Specifically, \mathbf{X} is first fed to the feature extractor. To address distribution shifts [43, 44], we introduce Instance Normalization (IN) [43]. The improved Positional Embedding (PE) is then used to incorporate temporal positional encoding, enabling SimAD to learn the relationships between C channels actively. The improved positional encoding (PE) applies sinusoidal positional encoding solely to the temporal positions, in contrast to the original PE [45], without the necessity of encoding the channel indices. This distinction arises from the differing implications of the channels in time series and the word embeddings in the language model.

Then, for C channels, the temporal sequence is processed through two operations via $\text{Patching}(\cdot)$. First, the original input $\mathbf{X} \in \mathbb{R}^{T \times C}$ is segmented into M patches of length P , resulting

in an intermediate representation $\mathbf{N}' \in \mathbb{R}^{M \times P \times C}$, where $T = M \times P$. Next, \mathbf{N}' is reshaped to obtain $\mathbf{N}' \in \mathbb{R}^{M \times (P \cdot C)}$. In this paper, we use the superscript “ r ” to denote intermediate process variables, such as \mathbf{N}' . Finally, Value Embedding (VE) is a simple linear transformation that maps all patches into a unified D -dimensional space, specifically $\mathbf{N} \in \mathbb{R}^{M \times D}$. It is defined as $\mathbf{N} = \text{VE}(\mathbf{N}') = \text{LayerNorm}(\text{Linear}(\text{LayerNorm}(\mathbf{N}')))$. Here, D is a predefined and represents the final dimension of the model. The term LayerNorm refers to layer normalization [45], which has been shown to be more suitable for sequential data. The term Linear refers to linear layer. The processing performed by the Feature Extractor can be described as below:

$$\begin{aligned} \mathbf{X}' &= \text{PE}(\text{IN}(\mathbf{X})), \\ \mathbf{N} &= \text{VE}(\text{Patching}(\mathbf{X}')), \end{aligned} \quad (1)$$

where \mathbf{N} refers to the **naive representation** and the original input of EmbedPatch Encoder. With the designed feature extractor that maintains the framework’s simplicity, SimAD can handle data with longer time windows, whose necessity will be demonstrated by in Section IV-C.

3) *EmbedPatch Encoder*: To enhance SimAD’s capability of modeling the dissimilarity between normal and abnormal samples, we made an improvement to the original attention mechanism. Inspired by [9, 11, 12], this improvement involves incorporating the EmbedPatch $\mathbf{E}^{(i)} \in \mathbb{R}^{V \times D}$ with embedded queries into the EmbedPatch Encoder backbone of SimAD, where V denotes the number of patch embeddings (EmbedPatch) and (i) for i -th layer of encoder. Determining a proper V often requires prior and a larger number of embeddings to accommodate different scenarios [46]. Therefore, a linear projection is stacked on EmbedPatch to obtain $\mathbf{E}^{(i),r} \in \mathbb{R}^{M^* \times D}$ for reducing the dimensionality, where $M^* \ll V$. Here, M^* represents the number of embeddings. It equals M from the Feature Extractor ($M^* = M$), but they serve different purposes in the context. Embedding (EmbedPatch) is learnable, whereas **naive representation** refers to the raw, high-

dimensional features extracted directly from the original time series. Intuitively, each of the M patches finds a corresponding embedding from the V patch embeddings, resulting in M^* new embeddings for the original patches. This simple component allows SimAD to efficiently learn key information from normal samples without being overly dependent on prior knowledge.

To achieve this, we employ an improved multi-head attention. For each head $k \in [1, 2, \dots, U]$, we define the query matrix as $\mathbf{Q}_k^{(i)} = \mathbf{N}^{(i)} \mathbf{W}_k^Q$, the key matrix as $\mathbf{K}_k^{(i)} = \mathbf{N}^{(i)} \mathbf{W}_k^K$, and the value matrix as $\mathbf{V}_k^{(i)} = \mathbf{E}_k^{(i),\prime} = \mathbf{W}_k^V \mathbf{E}_k^{(i)}$, where $\mathbf{W}_k^Q, \mathbf{W}_k^K \in \mathbb{R}^{D \times d}$, $\mathbf{W}_k^V \in \mathbb{R}^{M^* \times V}$, $\mathbf{E}^{(i)} = [\mathbf{E}_1^{(i)}, \mathbf{E}_2^{(i)}, \dots, \mathbf{E}_U^{(i)}]$, $\mathbf{E}_k^{(i)} \in \mathbb{R}^{V \times d}$, and d is defined as $\lfloor \frac{D}{U} \rfloor$, which represents the dimension of the head. For the first layer of the EmbedPatch Encoder, its input $\mathbf{N}^{(1)}$ is the output \mathbf{N} from the Feature Extractor. Next, we define the attention calculation between different patches in each layer as below:

$$\mathbf{Z}_k^{(i)} = \text{Attention}(\mathbf{Q}_k^{(i)}, \mathbf{K}_k^{(i)}, \mathbf{V}_k^{(i)}) = \text{Softmax}\left(\frac{(\mathbf{Q}_k^{(i)} \mathbf{K}_k^{(i)\top})}{\sqrt{d}}\right) \mathbf{V}_k^{(i)}. \quad (2)$$

Note that the $\mathbf{Q}_k^{(i)}$ and $\mathbf{K}_k^{(i)}$ here are generated by different parameters, so the attention scores are asymmetric. On the top of each layer, we utilize a linear layer to aggregate features $\mathbf{Z}_k^{(i)}$ from different heads, resulting in $\mathbf{Z}^{(i)} \in \mathbb{R}^{M \times D}$. Specifically, we first concatenate the features from U heads along the last dimension. For each $\mathbf{Z}_k^{(i)} \in \mathbb{R}^{M \times d}$, the concatenation yields $\mathbf{Z}^{(i),\prime} (\in \mathbb{R}^{M \times D}) = [\mathbf{Z}_1^{(i)}, \dots, \mathbf{Z}_U^{(i)}]$. Subsequently, this concatenated feature is aggregated via a linear layer Linear, resulting in $\mathbf{Z}^{(i)} = \text{Linear}(\mathbf{Z}^{(i),\prime})$.

Subsequently, we proceed with the original “Add & Norm” operation to generate the input, $\mathbf{N}^{(i+1)}$, for the next layer:

$$\begin{aligned} \mathbf{N}^{(i)\prime} &= \text{LayerNorm}(\mathbf{N}^{(i)} + \mathbf{Z}^{(i)}), \\ \mathbf{N}^{(i+1)} &= \text{LayerNorm}(\mathbf{N}^{(i)\prime} + \text{FFN}(\mathbf{N}^{(i)\prime})), \end{aligned} \quad (3)$$

where $\text{FFN}(\cdot)$ denotes a 2-layer feed-forward network (FFN), as in [45], the superscript \mathbf{N}' to denote intermediate variables, and the subscript (i) to indicate features at different layers.

Finally, a linear layer is used to restore $\hat{\mathbf{X}}$ from the last layer. As shown in Fig. 2(a), the final output of the EmbedPatch Encoder is processed by a linear layer, which can be expressed as $\hat{\mathbf{X}} = \text{Linear}(\mathbf{N}^{(-1)})$. Here, $\mathbf{N}^{(-1)}$ represents the features from the last layer of the encoder.

To enable the model to detect anomalies in long-context time series, we need to optimize the parameters of SimAD’s Feature Extractor and EmbedPatch Encoder. During the training phase, the Mean Square Error (MSE) [13, 19] is used to measure the difference between \mathbf{X} and $\hat{\mathbf{X}}$, and a patch-based similarity loss is employed to ensure continuity between patches. The MSE term ensures that the model reconstructs the time series accurately locally. However, this point-to-point reconstruction alone cannot guide the model to detect anomalies over long time windows. The similarity term addresses this limitation. It first converts \mathbf{X} into $\mathbf{N} \in \mathbb{R}^{M \times (P \cdot C)}$ using the previous Patching operation. Then, it calculates patch-based similarity along the last dimension. The final training loss is given by Eq. (4).

During the testing phase, SimAD utilizes \mathcal{L}_{rec} to calculate anomaly scores and detect anomalies.

$$\mathcal{L}_{rec} = \text{MSE}(\hat{\mathbf{X}}, \mathbf{X}) + (1 - \text{Similarity}(\hat{\mathbf{X}}, \mathbf{X})), \quad (4)$$

where cosine similarity is adopted as a similarity measure.

4) ContrastFusion Module: Although the EmbedPatch Encoder enables SimAD to memorize normal samples, it does not substantially enhance the discrimination between normal and abnormal data. To mitigate this limitation, the ContrastFusion module is designed to employ contrastive learning to enlarge the feature distance between normal and abnormal data’s features.

To generate negative samples for contrastive learning without introducing prior, we adopt the simplest method by using Gaussian noise [9, 1]. Specifically, we use the equation $\mathbf{X}^- = \mathbf{X} + \alpha \cdot \mathbf{J}$, where \mathbf{J} is sampled from a Gaussian distribution, and α controls the level of noise. This is a common time series data augmentation method [9, 1]. From then on, we use variables with subscript “-” to represent negative samples or features. Inspired by denoising autoencoders [47], we incorporate a denoising loss into the final objective function. Similarly, to guide the model to focus on contextual temporal associations during denoising, we added an extra similarity term.

$$\mathcal{L}_{denoise} = \text{MSE}(\hat{\mathbf{X}}^-, \mathbf{X}^+) + (1 - \text{Similarity}(\hat{\mathbf{X}}^-, \mathbf{X}^+)). \quad (5)$$

\mathbf{X}^+ indicates the branch opposite to the negative sample \mathbf{X}^- , emphasizing their opposition and being differentiated by colors/symbols in Fig. 2(a) for convenience and consistency. Keep in mind that $\mathbf{X}^+ = \mathbf{X}$ in this case. The two branches of ContrastFusion, which are optimized without regard to reconstruction, share structural similarities with previous image self-supervised techniques such as SimCLR [48] and SimSiam [49]. Therefore, our methodology can also be considered a self-supervised (and/or unsupervised [4]) learning strategy.

To learn invariant features, we feed positive samples \mathbf{X}^+ and negative samples \mathbf{X}^- into the Feature Extractor and EmbedPatch Encoder, resulting in the acquisition of representations \mathbf{N}^+ and \mathbf{N}^- . Previous self-supervised works [48, 49, 10, 1] have shown that constructing different views through data augmentation can guide models to learn invariant features. For this purpose, we utilize a projection head $\mathcal{P}(\cdot)$, akin to prior contrastive learning methods, to facilitate the acquisition of meaningful and discriminative representations.

$$\mathbf{H}^+ = \mathcal{P}(\mathbf{N}^+) = \text{Linear}(\text{ReLU}(\text{Linear}(\mathbf{N}^+))), \quad (6)$$

where ReLU is the activation function. We derive low-dimensional representations \mathbf{H}^+ and \mathbf{H}^- for \mathbf{N}^+ and \mathbf{N}^- respectively. To design an asymmetric feature contrastive loss, we utilize gradient stopping.

$$\begin{aligned} \mathcal{L}_{cont} &= \text{MSE}(\mathbf{H}^+, \text{StopGrad}(\mathbf{H}^-)) + (1 - \text{Similarity}(\mathbf{H}^+, \text{StopGrad}(\mathbf{H}^-))) \\ &\quad + \text{MSE}(\mathbf{H}^-, \text{StopGrad}(\mathbf{H}^+)) + (1 - \text{Similarity}(\mathbf{H}^-, \text{StopGrad}(\mathbf{H}^+))). \end{aligned} \quad (7)$$

5) Joint optimization for Training: As described above, SimAD aims to optimize the reconstruction of positive samples, denoise noisy samples, and amplify the feature differences between positive and negative samples (see in Fig. 2(b)). During the initial stages of training, the model focuses primarily on reconstruction and denoising. Since the contrastive loss may introduce additional training complexity, the weight of contrastive loss incrementally rises in the initial $N_{\text{warm-up}}$ iterations until it reaches the maximum value of β_{\max} : $\beta_i = \min\{\frac{i+1}{N_{\text{warm-up}}}, \beta_{\max}\}$.

The overall objective function combines the reconstruction loss, denoising loss, and contrastive loss, as below:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{denoise} - \beta \mathcal{L}_{cont}. \quad (8)$$

Algorithm 1 shows the training workflow of SimAD, highlighting the simplicity of our method in design and implementation.

Algorithm 1 The Pseudo-Code for Training SimAD.

```

1 # J: I.I.D GAUSSIAN NOISE
2 # FE: FEATURE EXTRACTOR
3 # ENC: EMBEDPATCH ENCODER
4 # CF: CONTRASTFUSION
5 # SIMAD: COMBINATION OF FE, ENC, AND
   CF
6 # R: REARRANGE DATA
7 for x in data_loader:
8     x1, x2 = x, x + J
9     x_out1, sim1 = SimAD(x1)    # SIMAD
        PROCESSING
10    x_out2, sim2 = SimAD(x2)
11    # RECONSTRUCTS AND PROJECTS
        FEATURES
12    x_patch = R(x)
13    rec_loss = loss_func(x_out1,
        x_patch)
14    denoise_loss = loss_func(x_out2,
        x_patch)
15    sim_loss = loss_func(sim1,sim2.
        detach()) + loss_func(sim2,sim1.
        detach())
16    loss = rec_loss + denoise_loss -
        sim_loss * warmup
17    loss.backward()
18    update(SimAD)
19
20 # LOSS FUNCTION
21 def loss_func(s1,s2):
22     return 12_loss(s1,s2) + (1-
        cos_loss(s1,s2)).mean()

```

B. Baselines

1) *Inference: Anomaly score:* To calculate the anomaly score for query samples, we use the following equation:

$$\mathcal{L}_{score} = \text{MSE}(\hat{\mathbf{X}}, \mathbf{X}) + (1 - \text{Similarity}(\hat{\mathbf{X}}, \mathbf{X})). \quad (9)$$

Note that the first term (MSE) has a length of T time points, while the second term (similarity) has a length of M patches. To match the length, each patch in the similarity term is replicated P times to reach T time points in the implementation.

As shown in Fig. 2(c), this process involves inputting time series data into SimAD and performing feature extraction and representation to obtain the final feature $\mathbf{Z}^{(L)}$. This feature is then fed into a linear layer to reconstruct the data as closely as possible to the original: $\hat{\mathbf{X}} = \text{Linear}(\mathbf{Z}^{(L)})$. We then compute the MSE and cosine similarity between the reconstructed and

original data. In inference stage, the ContrastFusion module is not used, eliminating the need to generate negative samples or compute similarities between positive and negative samples.

C. Improving Affiliation Metrics

Recent studies [15, 16, 17] have highlighted the limitations of conventional metrics used in time series anomaly detection. The affiliation metric [50] is a parameter-free method that has shown promising performance. However, based on experiments and theoretical analysis (see Appendix C), the Affiliation precision (Aff-Pre) tends to approach 0.5, while the Affiliation recall (Aff-Rec) tends to approach 0.99 when confronting with random anomaly scores. This leads to an Affiliation F1 score (Aff-F1) of approximately 0.7, suggesting that the metric's discriminatory capability is insufficient. To address this limitation, we introduce **Unbiased Affiliation (UAff)** and **Normalized Affiliation (NAff)** metrics. Both of them exhibit superior discriminatory ability and provide a more accurate reflection of an algorithm's performance.

Experimental findings (Table I) indicate that the Affiliation precision (Aff-Pre) varies across different datasets but consistently tends to approach 0.5, which can be interpreted as a dataset-specific bias, denoted by Aff-Pre_{bias} . Meanwhile, it is noteworthy that there has been insufficient encouragement for high-precision models, while the penalties for low-precision models are inadequate. In this context, we design Unbiased Affiliation precision (UAff-Pre) and Unbiased Affiliation F1 (UAff-F1), both of which can alleviate the dataset-specific bias and offer a more equitable assessment of the algorithm's precision performance.

$$\text{UAff-Pre} = \frac{\text{Aff-Pre} - \text{Aff-Pre}_{bias}}{1 - \text{Aff-Pre}_{bias}} \quad (10)$$

$$\text{UAff-F1} = \frac{2 \cdot |\text{UAff-Pre}| \cdot \text{Aff-Rec}}{|\text{UAff-Pre}| + \text{Aff-Rec}} \cdot (-1)^{|\text{UAff-Pre}| < 0}. \quad (11)$$

where **UAff-F1** is negative when **UAff-Pre** < 0 . Although **UAff-Pre** denotes an improvement over its original version and surpasses **UAff-F1**, it does introduce additional parameters, thereby deviating from its parameter-free nature. Since the majority of real-world datasets exhibit $\text{Aff-Pre}_{bias} \leq 0.55$, we further propose a modified version of NAff metrics. The computation follows a similar process, with the exception that we set $\text{Aff-Pre}_{bias} = \beta$ for constant while maintaining other aspects unchanged. In other words, **NAff-Pre** can be calculated as $\frac{\text{Aff-Pre}-\beta}{1-\beta}$, and **NAff-F1** is calculated similarly to **UAff-F1**. The mathematical and experimental analyses of the metrics are in the Sec. C. C.

IV. EXPERIMENTS

We evaluate the effectiveness of our proposed SimAD by conducting extensive comparison experiments against state-of-the-art competing methods on six real-world multivariable datasets: MSL, SMAP, PSM, SWaT, WADI, and Swan. Additionally, we utilize a univariate dataset UCR [51]. The details of the above six multivariable datasets are in Table I.

We employ the F1 without point adjustment, Aff-F1, and the improved UAff-F1 and NAff-F1 for performance evaluation.

TABLE I
DETAILS OF BENCHMARK DATASETS.

Dataset	#Training	#Test (Labeled)	Dimension	Anomaly ratio (%)	Bias	Ideal Bias
MSL	58317	73729	55	10.5	51.34	50.55
SMAP	135183	427617	25	12.8	51.48	50.82
SWaT	99000	89984	26	12.2	52.94	50.74
WADI	241921	34561	123	5.74	54.81	50.16
PSM	132481	87841	25	27.8	53.17	53.86
Swan	60000	60000	38	32.6	54.83	55.31

We exclude the point-adjusted F1 score due to its acknowledged potential for false improvements. It is noted that although the classical F1 score is not optimal for time series data, it can still be employed for evaluation purposes [15], as detailed in Appendix D-A. Furthermore, we present an analysis of VUS [52] scores (Table X in Appendix) for all the datasets. In TSAD, the F1 score evaluates model performance by balancing precision (Prec) and recall (Rec). It is defined as:

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad \text{F1} = 2 \times \frac{\text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}. \quad (12)$$

Here, TP, FP, and FN represent true positives, false positives, and false negatives. The F1 score ranges from 0 to 1, with higher values indicating better anomaly detection performance. Due to space limitations and the need for conciseness, the calculation of Aff-F1 and VUS refer to the original papers [50, 52].

We extensively discuss UAFF and NAFF in Sec. C, comparing them with other metrics, and summarize the limitations and advantages of existing metrics in detail.

A. Baselines

Our model is compared against 20 baselines, which include machine learning-based models like LOF, Isolation Forest (IForest) and PCA; deep learning-enhanced models like Deep SVDD [26] and Deep Isolation Forest (Deep IF) [27]; reconstruction-based models like USAD [19], TCN-ED [14], AdaMemBLS [1] and NPSR [53]; prediction-based methods like TimesNet [30], and M2N2 [54]; Transformer-based models such as Anomaly Transformer (AnomTrans) [13] and TranAD [21]; contrastive learning-based models like COUTA [35], NCAD [36], and DCdetector [20]; diffusion-based model D3R [39]; and LLM-based GPT2-Adapter [41].

B. Quantified Comparisons

Table II shows comparison results between our proposed SimAD and baselines on six real-world datasets. It is obvious that SimAD performs superior to other models, particularly on datasets SWaT, WADI, and Swan. Specifically, compared to the best baseline on each dataset, SimAD exhibits an improvements of **5.15%** (76.88%→82.03%), **13.31%** (50.67%→63.98%), and **15.24%** (55.99%→71.23%) in F1 score. In contrast, the NAFF-F1 score exhibits absolute improvements of **11.09%** (55.73%→66.82%), **30.87%** (45.39%→76.26%), and **19.05%** (33.99%→53.04%) respectively. In addition, SimAD has shown slight superiority over other models on datasets MSL and SMAP. On these datasets, SimAD achieved an absolute improvement

of 1.26% and 1.00% in Aff-F1, respectively, along with higher scores in other metrics as well.

Table III summarizes the evaluation metric scores of models across all datasets, clearly indicating that SimAD achieved an improvements of **9.07%** on F1, **3.18%** on Aff-F1, **16.63%** on UAFF-F1, and **20.55%** on NAFF-F1, and **6.83%** AUC, compared to the SOTA baseline. In contrast, although models such as TimesNet and D3R exhibit good performance on specific evaluation metrics (UAFF-F1, Aff-F1), their performances on other metrics (Aff-F1, NAFF-F1) are inferior to those of SimAD, and even lower than those of random algorithms. Fig. 3 shows the advantages of our model, compared to other baseline models across multiple evaluation metrics. This further confirms the superior generalization capability of SimAD, whereas other algorithms may display larger performance fluctuations due to an excessive focus on specific dataset characteristics. (See a more detailed comparison in Appendix E-D.)

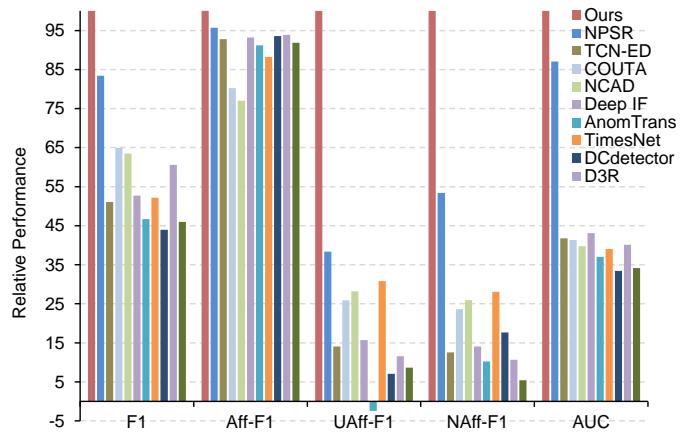


Fig. 3. Comparison analysis of relative performance.

Additionally, we evaluate the effectiveness of our model in handling univariate dataset UCR in Table IV, which includes Avg.+ (calculated using only positive values) and Avg. scores. From this table, SimAD demonstrates superior performance. Specifically, SimAD achieves an improvement of **13.33%** (1.66%→14.99%) in **Avg.+ F1** and **13.38%** (1.61%→14.99%) in **Avg. F1**, compared to the best baseline. Furthermore, SimAD achieves an improvement of **25.75%** (9.41%→35.16%) in **Avg.+ UAFF-F1** and **17.52%** (2.01%→19.53%) in **Avg. UAFF-F1**, compared to the best baseline.

C. Sensitivity Analysis

We performed sensitivity analysis on hyperparameters of SimAD, which include the window size and the number of patch embeddings. Fig. 4 illustrates the influence of different window sizes on the model's performance on the WADI dataset. As the window size increases, the model's loss exhibits a decreasing trend, and its accuracy consistently improves. A larger window size incorporates more temporal and contextual information, enabling the model to capture patterns and trends and improve the detection performance. It is worth noting that when the window size increases from 1024 to 2048, the model's loss further decreases while the improvement in the

TABLE II
COMPARISON RESULTS. ALL RESULTS ARE IN %, THE BEST IN **BOLD**, AND THE SECOND IN UNDERLINE.

Datasets	MSL										SMAP										SWaT									
Methods	F1	AUC	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1	F1	AUC	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1	F1	AUC	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1									
Random	18.60	50.01	51.34	99.99	67.84	0.00	5.20	22.06	50.18	51.48	100.0	67.97	0.00	5.74	21.06	50.10	52.94	99.99	69.23	0.00	11.09									
LOF	19.36	55.75	53.57	88.90	66.86	8.74	13.22	23.64	62.39	46.00	81.26	58.74	-19.84	-14.58	58.22	84.59	99.87	2.70	5.25	5.25	5.25									
IForest	10.90	59.09	55.09	97.63	70.44	14.31	18.45	23.78	61.15	39.45	70.23	50.52	-36.64	-32.45	38.52	83.80	100.0	2.71	5.27	5.27	5.27									
PCA	10.33	53.25	55.50	97.44	70.72	15.73	19.77	22.96	58.70	39.67	70.31	50.72	-36.15	-31.93	26.05	81.85	99.96	2.70	5.25	5.25	5.25									
Deep SVDD	24.30	61.80	52.67	99.93	68.98	10.53	10.13	21.28	61.14	44.98	88.45	59.64	-17.60	-18.02	22.57	86.81	53.58	<u>99.99</u>	69.77	13.75	13.37									
USAD	22.50	57.12	51.90	97.29	67.69	7.73	7.31	16.57	62.79	39.56	69.48	50.41	-31.78	-32.12	21.70	88.66	53.00	<u>100.0</u>	69.28	11.71	11.31									
TCN-ED	19.66	51.14	51.44	99.99	67.93	6.04	5.61	22.90	58.91	51.39	67.89	5.86	5.42	21.65	89.23	52.98	100.0	69.26	11.64	11.24										
COUTA	20.88	55.59	51.24	99.65	67.68	5.29	4.85	22.69	58.74	51.48	100.0	67.97	6.18	5.74	47.65	75.54	79.98	33.20	46.92	42.76	42.73									
TranAD	22.81	50.03	52.82	99.34	68.97	11.07	10.66	22.74	59.74	39.41	70.30	50.51	-32.21	-32.56	25.50	88.90	55.65	99.66	71.42	20.64	20.30									
NCAD	22.05	60.20	56.38	83.30	67.25	22.46	22.14	23.09	53.45	51.88	<u>99.99</u>	68.32	7.67	7.25	68.72	82.92	65.01	85.58	73.89	44.64	44.46									
Deep IF	19.11	55.94	51.45	100.0	67.94	6.08	5.64	29.14	60.09	53.75	98.67	69.59	14.31	13.93	21.65	89.52	52.98	100.0	69.26	11.64	11.24									
AnomTrans	18.39	52.61	50.21	99.83	66.81	-4.53	0.83	16.06	52.18	55.84	99.11	71.44	16.49	20.90	23.44	80.80	50.24	99.49	66.77	-10.83	0.96									
TimesNet	21.24	57.18	51.12	99.95	67.64	4.81	4.36	24.37	53.73	49.50	99.86	66.19	-1.52	-1.99	21.66	73.24	57.84	93.77	71.55	27.16	26.86									
DCdetector	11.62	50.31	51.96	97.77	67.85	2.52	7.52	26.56	58.50	55.55	99.78	71.37	15.48	19.97	23.24	52.78	52.63	98.30	68.56	-1.28	10.00									
D3R	23.98	63.00	53.61	99.97	69.79	8.92	7.65	22.71	54.56	51.43	100.0	67.92	-0.21	0.07	45.89	79.95	61.47	78.52	68.96	29.47	36.02									
GPT2-Adapter	13.72	52.03	53.31	97.01	68.81	7.80	6.75	24.12	55.48	53.28	99.84	69.48	7.18	3.67	22.30	52.30	52.51	98.13	68.41	-1.79	0.07									
NPSR	23.72	61.16	52.05	99.81	68.42	2.89	7.88	22.68	61.26	51.46	100.0	67.95	-0.06	5.68	76.88	90.18	71.21	81.16	<u>75.86</u>	<u>52.54</u>	<u>55.73</u>									
M2N2	21.76	59.15	51.38	100.00	67.88	0.17	5.36	22.68	54.05	51.46	100.00	67.95	0.52	5.69	38.50	67.79	59.15	96.83	73.44	27.54	30.77									
AdaMemBLS	19.11	51.78	57.07	95.37	<u>71.41</u>	<u>20.99</u>	24.64	26.93	53.66	53.30	99.80	69.49	7.24	12.38	74.10	81.61	53.57	100.00	69.76	2.65	13.32									
Ours	30.02	62.70	56.30	99.76	71.98	18.50	<u>22.37</u>	29.39	65.46	56.84	99.82	72.44	19.91	24.07	82.03	90.31	78.46	80.88	79.65	64.93	66.82									
Datasets	WADI										PSM										NIPS-TS-Swan									
Methods	F1	AUC	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1	F1	AUC	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1	F1	AUC	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1									
Random	10.79	50.03	54.81	99.95	70.79	0.00	17.54	40.93	50.16	53.17	99.95	69.41	0.00	11.91	46.25	50.13	54.83	96.82	70.01	0.00	17.56									
LOF	10.58	65.44	51.76	25.72	34.36	-10.67	6.20	23.91	81.28	77.47	52.13	62.32	52.01	53.50	21.37	77.89	82.01	11.14	19.62	18.81	18.98									
IForest	30.88	74.84	57.68	82.53	67.90	11.82	25.91	38.52	75.12	72.97	46.46	56.77	44.27	46.20	46.18	85.12	94.11	16.31	27.81	27.47	27.54									
PCA	34.30	50.53	48.10	52.07	50.01	-23.09	-7.08	43.20	72.06	77.86	78.74	63.44	65.57	49.18	72.88	92.96	16.03	27.35	26.95	27.02										
Deep SVDD	7.34	59.47	50.06	55.79	52.77	0.72	0.24	44.46	78.09	53.92	<u>99.98</u>	70.06	14.93	14.56	<u>55.99</u>	75.32	56.84	89.93	69.66	24.07	23.76									
USAD	10.86	53.61	54.92	100.0	70.90	18.27	17.91	47.90	64.53	55.30	98.71	70.88	19.48	19.14	52.63	66.76	55.26	68.62	61.22	18.57	18.25									
TCN-ED	10.86	53.01	54.92	100.0	70.90	18.27	17.91	43.46	63.97	53.16	100.0	69.42	12.29	11.89	49.22	71.37	55.10	99.92	71.03	18.87	18.52									
COUTA	26.92	50.52	99.18	23.08	37.44	<u>37.38</u>	37.38	48.21	69.36	57.98	99.30	73.21	27.80	27.50	47.19	71.38	45.07	86.50	66.54	15.24	14.87									
TranAD	11.78	52.60	55.91	91.96	69.54	21.27	20.94	43.20	65.22	60.88	92.09	73.30	35.44	35.19	51.93	70.00	57.86	90.18	70.49	27.07	26.78									
NCAD	10.87	62.84	54.99	100.0	70.96	18.49	18.14	46.61	61.76	77.79	53.08	63.10	54.35	54.30	37.57	52.24	54.32	94.06	68.87	16.20	15.83									
Deep IF	10.86	51.43	54.92	100.0	70.90	18.27	17.91	43.47	69.06	53.15	100.0	69.41	12.24	11.84	49.19	73.30	55.10	99.98	71.05	18.86	18.51									
AnomTrans	11.24	52.98	51.43	95.15	66.77	-13.84	5.56	39.81	52.18	52.78	96.07	68.13	-1.65	10.51	44.86	54.62	55.21	92.36	69.11	1.68	18.73									
TimesNet	17.65	65.08	65.25	39.30	49.05	34.45	34.34	43.60	58.74	77.84	67.23	72.15	69.97	60.91	43.16	53.62	60.75	81.29	69.53	<u>34.23</u>	<u>33.99</u>									
DCdetector	11.33	50.12	59.72	94.64	73.23	19.51	32.26	22.72	50.38	53.05	94.98	68.08	-0.49	11.47	48.83	50.35	55.08	99.82	70.99	1.11	18.45									
D3R	12.88	51.39	56.47	<u>99.98</u>	72.17	7.09	2.46	44.24	60.64	53.84	100.0	69.99	2.82	3.93	49.45	62.54	57.70	96.06	<u>72.10</u>	11.93	9.98									
GPT2-Adapter	10.68	51.21	55.62	<u>98.29</u>	71.04	3.55	-0.27	35.24	51.23	54.33	95.16	69.17	4.84	2.18	45.23	58.77	61.17	69.55	<u>65.09</u>	23.36	18.28									
NPSR	50.67	80.19	65.15	90.45	75.74	36.53	<u>45.39</u>	<u>51.02</u>	70.67	54.71	99.37	70.56	6.29	17.14	49.30	62.61	55.13	99.99	71.07	1.32	18.61									
M2N2	13.55	55.68	54.92	100.00	70.90	13.73	17.93	47.71	54.91	55.43	94.36	69.84	15.45	19.48	38.40	70.12	60.94	93.68	73.84	32.60	35.47									
AdaMemBLS	29.09	53.85	54.95	100.00	70.93	0.63	18.01	46.69	63.26	58.22	71.51	19.32	27.92	49.35	51.27	55.14	100.00	71.09	1.38	18.66										
Ours	63.98	87.97	82.36	92.81	87.27	73.59	76.26	52.07	71.99	62.46	91.76	74.33	32.63	39.20	71.23	85.17	76.17	53.75	63.03	50.29	53.04									

TABLE III
THE AVERAGE PERFORMANCE AND RANKING OF DIFFERENT ALGORITHMS.
RK. DENOTES THE RANKING.

Avg.	F1	RK.	Aff-F1	RK.	UAff-F1	RK.	NAff-F1	RK.	AUC	RK.	V_PR	RK.	Avg.	RK.	
Random	26.62	16	69.21	9	0.00	19	11.51	15	50.10	20	27.71	18	16.17		
LOF	26.18	17	41.19	20	9.05	13	13.76	10	71.23	3	39.75	9	12.00		
IForest	31.46	7	46.45	19	11.09	11	15.15	9	73.19	2	43.39	7	9.1		

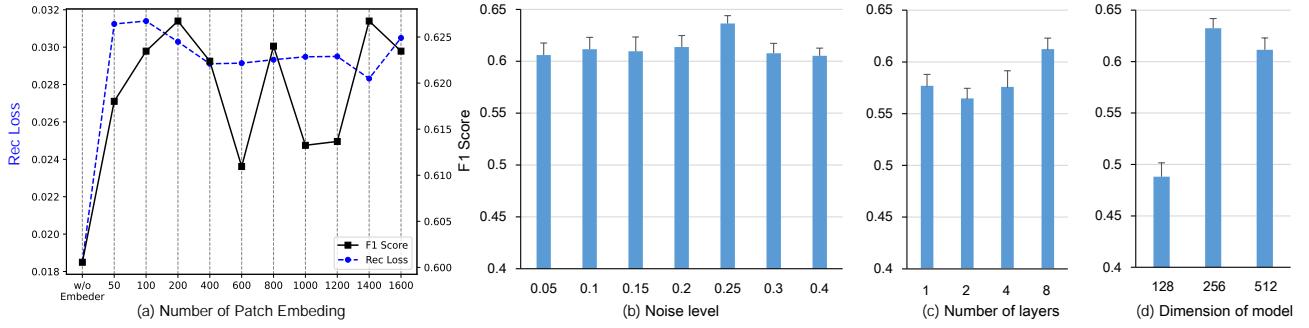


Fig. 5. Different numbers of embeddings and other hyper-parameters. The error bar represents the standard deviation.

TABLE V
FULL ABLATION RESULTS.

Datasets	MSL						SMAP						SWaT					
Variations	F1	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1	F1	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1	F1	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1
w/o Cont.	23.90	54.13	99.08	70.01	10.84	15.23	22.56	51.94	99.86	68.34	1.89	7.47	80.31	77.66	76.52	77.09	62.29	64.21
w/o Aug	29.07	54.64	98.88	70.39	12.72	16.98	23.78	52.66	98.11	68.53	4.75	10.09	79.14	80.60	64.14	71.43	61.34	62.63
w/o Embedder	27.20	54.12	98.85	69.94	10.80	15.20	17.89	50.04	99.64	66.62	-5.76	0.15	80.31	77.66	76.52	77.09	62.29	64.21
w/o Asym.	28.23	54.13	98.94	69.98	10.86	15.25	22.39	51.46	99.77	67.90	-0.05	5.69	82.01	77.35	83.37	80.25	63.95	66.06
w/o Cos	28.50	53.95	99.10	69.87	10.20	14.64	23.32	52.24	99.91	68.61	3.11	8.59	77.77	75.02	83.53	79.05	60.09	62.59
Ours	30.02	56.30	99.76	71.98	18.50	22.37	29.39	56.84	99.82	72.44	19.91	24.07	82.03	78.46	80.88	79.65	64.93	66.82
Datasets	WADI						PSM						NIPS-TS-Swan					
Variations	F1	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1	F1	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1	F1	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1
w/o Cont.	62.68	70.55	84.04	76.71	49.26	55.20	45.78	58.83	96.05	72.96	21.47	29.83	64.34	70.24	57.00	62.93	42.69	47.34
w/o Aug	63.19	75.92	81.53	78.63	59.40	63.38	41.79	56.43	96.15	71.12	12.98	22.67	64.83	62.55	65.81	64.14	27.13	36.34
w/o Embedder	61.36	70.73	59.61	64.70	44.29	48.91	49.78	61.47	94.80	74.58	29.86	36.93	68.32	71.04	62.03	66.23	45.47	50.14
w/o Asym.	56.89	63.33	63.98	63.65	56.26	37.64	52.26	59.93	91.27	72.35	24.93	32.62	69.14	68.85	40.28	50.82	35.06	38.95
w/o Cos	63.13	72.66	74.79	73.71	51.69	56.43	44.00	59.83	86.97	70.89	24.46	32.07	70.79	78.76	42.99	55.62	47.47	49.21
Ours	63.98	82.36	92.81	87.27	73.59	76.26	52.07	62.46	91.76	74.33	32.63	39.20	71.23	76.17	53.75	63.03	50.29	53.04

patch embedding is set to 0 (i.e., the original attention mechanism), the model’s performance is significantly lower. This demonstrates the effectiveness of EmbedPatch in enhancing the memory of normal samples. As the number of patch embedding increases, both the model’s Rec Loss and F1 Score exhibit fluctuating changes. It can be inferred that the model achieves the optimal performance when the number of patch embedding is between 100 and 200. From Fig. 5(b-d), it is evident that increasing the number of layers and setting an appropriate level of noise and dimension of features can result in improved model performance on the WADI dataset. However, a notable decline in performance is observed when the dimension is set to 128. This indicates that models with smaller dimensions face challenges in learning complex features.

D. Ablation Analysis

To validate the effectiveness of our model, we conducted ablation studies on ContrastFusion, data augmentation, EmbedPatch, and asymmetric optimization in Fig. 6. Here we provide the descriptions about the model’s variants:

1. **w/o Contrastive** (Cont.): indicates the removal of contrastive learning in the ContrastFusion. By removing this, \mathcal{L}_{cont} (Eq. (7)) no longer works, and the similarity relationships between positive and negative samples will not be learned.

2. **w/o Aug**: signifies negative sample generation without denoising learning, where $\mathcal{L}_{denoise}$ (Eq. (5)) does not work. It is noteworthy that despite generating negative samples, it is

still possible to conduct contrastive learning between positive and negative samples, i.e., \mathcal{L}_{cont} (Eq. (7)) remains active.

3. **w/o Embedder**: when the Embedder, i.e., \mathbf{E} , is removed, the architecture of SimAD regresses to a Transformer. Originally, the value matrix in SimAD was $\mathbf{V} = \mathbf{W}^V \mathbf{E}$. After regressing to a Transformer, the generation of \mathbf{V} aligns with matrices \mathbf{Q} and \mathbf{K} , i.e., $\mathbf{V} = \mathbf{N} \mathbf{W}^V$. For simplicity, \mathbf{W}^V represents learnable parameters, and \mathbf{N} is the pre-input. When EmbedPatch is removed, SimAD’s backbone reverts to a Transformer, where self-attention is uniform across \mathbf{Q} , \mathbf{K} , and \mathbf{V} .

4. **w/o Asymmetric** (Asym.): our asymmetric optimization relies on gradient stop. Removing the asymmetric optimization eliminates the StopGrad from the equation. As both MSE and Similarity losses are symmetric, after removing this operation, Eq. (7) regresses to a symmetric optimization method.

5. **w/o Cos**: indicates the removal of the cosine terms from Eq. (4), (5), and (7) while keeping the MSE loss.

The detailed ablation studies are in Table V. The observation from ablation studies are summarized as below:

- After removing the **w/o Contrastive**, i.e., the removal of positive and negative sample contrastive learning, the F1 score of the model decreased by 7.88%, indicating that ContrastFusion is important in learning distinct features to differentiate between normal and abnormal samples.
- By disabling **w/o Aug**, i.e., the removal of data augmentation, the detection performance of SimAD shows a decrease of 43.69% in UAff-F1, demonstrating its ability to improve the model’s generalization.

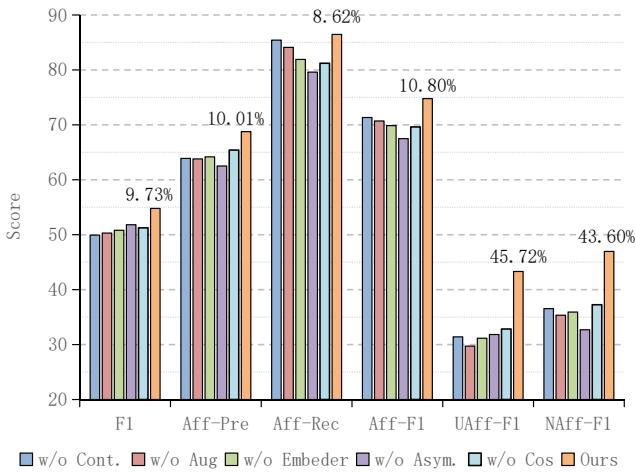


Fig. 6. Ablation study. The data labeling denotes an upper bound on the performance improvement of our model.

- 3) For **w/o Embeder**, the EmbedPatch in the attention mechanism degrades to a simple value matrix. The ablation of this module has the most significant impact on the model's F1 score, which indicates that incorporating EmbedPatch into the attention mechanism is crucial to learning appropriate representations of temporal data.
- 4) **w/o Asymmetric** indicates the model without asymmetric optimization. The ablation of this module results in a decrease in all performance metrics, particularly a significant drop of 43.69% in NAff-F1. This suggests that asymmetric optimization helps the model focus on abnormal samples.
- 5) The experimental results of **w/o Cos** demonstrate that upon removing the cosine similarity penalty, both the predictive continuity and detection performance of SimAD decline.

The comprehensive results of ablation studies, along with the setting of the experiment, are in Table V.

To verify the effectiveness of the EmbedPatch encoder and ContrastFusion modules in our proposed SimAD framework for addressing the first Challenge outlined in the introduction, *i.e.*, enhancing the dissimilarity between features of both normal and abnormal data, we extracted feature representations before and after the ContrastFusion projection head. Next, we conducted a multi-sampling to calculate the KL divergence between normal and abnormal features at both stages. The results are in Table VI. From this table, it can be observed that removing the ContrastFusion module (w/o Cont.) leads to an increase in KL divergence between the feature distributions of normal and abnormal data on the SWaT and WADI datasets, indicating that the representations become more distinct. Moreover, when both components are integrated into our model, the separation between normal and abnormal feature distributions is maximized. This demonstrates that SimAD effectively overcomes the limitations of previous approaches.

Moreover, Fig. 7 provides an illustration of the training on dataset PSM. It is evident that the **w/o Embedder** module achieves the lowest loss but with the lowest F1 score. This

TABLE VI
THE KL DIVERGENCE BETWEEN NORMAL AND ANOMALOUS FEATURES BOTH BEFORE AND AFTER PROJECTION.

Datasets	SWaT		WADI		
	Variations	Before Proj.	After Proj.	Before Proj.	After Proj.
w/o Cont.		7.34E-01	1.66E-02	4.95E-05	2.85E-03
w/o Embeder		4.34E-01	2.04E-02	3.25E-03	1.59E-01
Ours		4.25E-01	5.04E-01	3.00E-02	9.89E-01

can be attributed to that EmbedPatch contributes SimAD to memorizing normal samples and enhancing generalization. By capturing the characteristics of normal data, EmbedPatch improves the effectiveness of anomaly detection by leveraging the distinct differences between normal and abnormal samples, as demonstrated in Fig. 8(b) and Fig. 12 in Appendix E. The **w/o Contrastive** module also exhibits a low loss but suffers from negative optimization, resulting in a relatively low F1 score. In contrast, SimAD, despite having a higher loss, achieves the highest F1 score. This demonstrates that optimizing solely for reconstruction is inadequate, and SimAD effectively addresses the challenges outlined in the introduction.

E. Visualization Analysis

1) *Real-world cases analysis:* The performance of SimAD on dataset SWaT is in Fig. 8a and Fig. 10 (in Appendix E-B). The first row of images showcases the original data and the ground truth labels. Its y-axis represents the magnitude of the first channel in the time series. The second row of images denotes the predicted results, which effectively detects most anomalies by assigning significantly higher scores to abnormal data. The y-axis represents the magnitude of the anomaly score. The third and fourth rows of images illustrate the results obtained using L2 loss and patch-based cosine similarity loss as anomaly scores respectively.

It is noticed that the former focuses more on short-duration anomalies, while the latter can capture long-duration anomalies. In Appendix E-B, Fig. 11 shows the detection results of AnomTrans and NPSR. Our method exhibits superior performance compared to AnomTrans, as the latter only detects a few anomalies and encounters difficulties in accurately detecting segment anomalies. The case study highlights the effectiveness and rationale behind our anomaly score design.

2) *Explanation of patch embedding:* Fig. 8b shows the detection performance of SimAD, where the fourth row denotes the similarity between the query and the embedding of the last layer, while the vertical axes of the third and fourth regions illustrates correlations between the query embedding and the learned embedding of value matrices. In the third and the fourth regions of the right subfigure, the horizontal axis denotes the ordinal number of the embedding. The colors in the heatmap indicate the strength of these correlations, with precise values corresponding to the color bar for reference. Specifically, given a time window of 2048, we divide it into 64 non-overlapping patches, each with a length of 32. There are 64 learned embedding in the last layer of our EmbedPatch Encoder. In the following, we consider the 64 non-overlapping

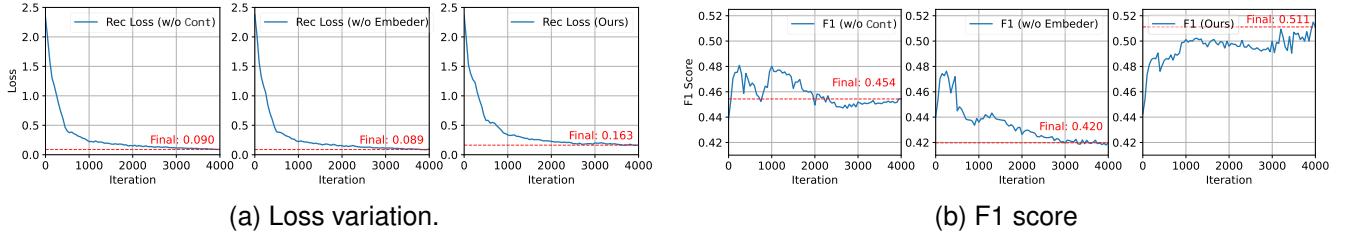


Fig. 7. Loss and F1 score variations on dataset PSM.

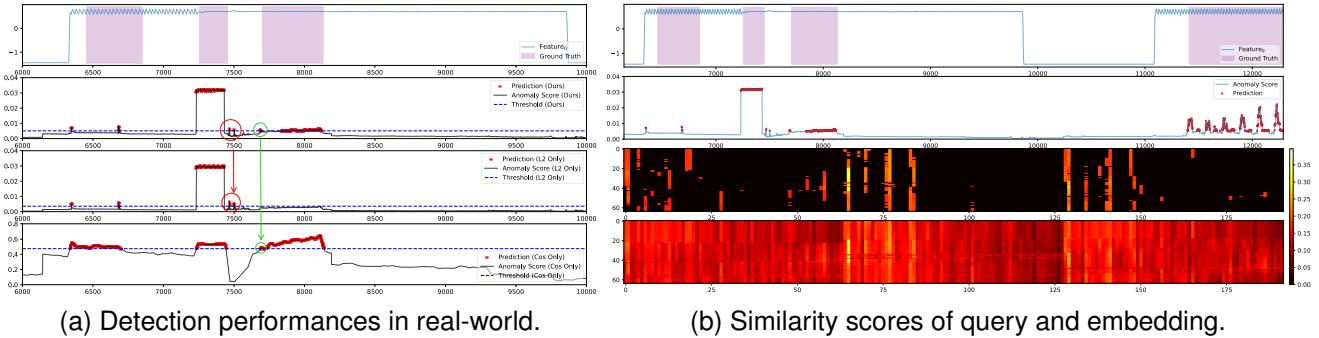


Fig. 8. Real world case study on dataset SWaT.

patches from the input window as the query embedding. Correspondingly, the 64 learned embedding produced by the final layer of the EmbedPatch Encoder serves as the value matrices. Next, we compute the pairwise correlations between the 64 non-overlapping patches and 64 learned embedding representations, resulting into a attention feature map. This feature map is depicted in the fourth region of the right subfigure. In contrast, the third region of the right subfigure highlights the top five most relevant query embeddings, selected from the 64 patches (i.e., segments of the 2048-length time window), for each learned embedding. In the third region, patches that exhibit weak relevance to all embeddings are shown in black. Conversely, patches with stronger relevance to certain embeddings are displayed in vibrant colors, highlighting their higher correlations relative to other patches. It is important to note that the attention feature map denotes the correlations between the query (\mathbf{Q}) vectors and the value (\mathbf{V}) vectors in the last layer of the EmbedPatch Encoder, i.e., the relationships between the input patches from the time series and the learned embedding, rather than the correlations between the query (\mathbf{Q}) vectors and the key (\mathbf{K}) vectors. Combined with Fig. 12 in Appendix E-B, it can be observed that in the lower layers, the embedding establishes relationships with a majority of queries. However, only the normal data can establish relationships with EmbedPatch in the higher layers. This means that in the shallow layers, SimAD learns general representations. While in the higher layers, SimAD learns features associated with normal data, which results in less similarity between abnormal data and EmbedPatch. Consequently, the patch embeddings enhance the detection performance of SimAD.

V. CONCLUSION

This paper introduces a **Simple** dissimilarity-based approach for time series Anomaly Detection (**SimAD**). The proposed framework is distinguished by its simplicity and versatility, allowing for the use of longer time windows. Experimental results demonstrate the superiority of our method compared to existing approaches. Additionally, we present two improved evaluation metrics, **UAff** and **NAff**, which effectively assess the algorithm's performance and overcome many shortcomings of previous metrics. In future work, we plan to extend this detection scheme to various other scenarios and aim to develop a unified model for anomaly detection.

ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China No. 62476101, 92467109, U21A20478, 62306052, National Key R&D Program of China 2023YFA1011601, the Major Key Project of PCL, China under Grant PCL2023AS7-1, and Natural Science Foundation of Chongqing under Grant CSTB2023NSCQ-LZX0092.

APPENDIX

This is the appendix of SimAD: A Simple Dissimilarity-based Approach for Time Series Anomaly Detection.

A. RELATED WORK

Unsupervised methods have garnered considerable attention due to the scarcity of labeled data for anomaly detection in time series sequences. These methods are particularly advantageous in scenarios where obtaining labeled data is expensive, time-consuming, or impractical. The primary objective is to identify outliers or anomalies that deviate significantly from expected

patterns without relying on prior annotations. Broadly, these methods can be categorized as follows:

(1) Algorithms based on classical machine learning, such as those proposed in [26, 27], transform traditional machine learning techniques into deep neural network architectures. This transformation enhances their ability to handle complex and high-dimensional data, allowing for more nuanced feature extraction and representation learning. By integrating classical algorithms with deep learning, these methods can capture intricate patterns that signify anomalies more effectively than their purely classical counterparts.

(2) Reconstruction-based methods, exemplified by [19, 14, 29], primarily involve training models on normal data to learn how to reconstruct input time series. During testing, the reconstruction error is used as an anomaly score, with anomalous data typically yielding higher errors due to their deviation from learned normal patterns. This approach relies on the assumption that normal data can be accurately reconstructed, while anomalies will result in significant discrepancies. For instance, USAD [19] employs a dual-decoder network structure and utilizes two-stage training to enhance performance. In this architecture, one decoder reconstructs the input data to capture normal patterns, while the second decoder focuses on learning the residuals—discrepancies that occur during reconstruction. This dual-decoder approach enables the model to effectively differentiate between normal and anomalous instances by analyzing both reconstruction fidelity and residual errors, thereby improving anomaly detection accuracy.

(3) Prediction-based techniques, as demonstrated by [30], focus on learning from historical data to predict future observations. These models are trained to forecast the next time step or series of steps based on past values. The prediction errors—defined as the difference between predicted and actual values—serve as the basis for anomaly detection. This method effectively identifies anomalies that manifest as significant deviations from expected future behavior. For example, PatchTST [55] employs a single-scale patch-based MLP as its core architecture, whereas the work in [56] applies an MLP to capture frequency-domain information in time series data for predictive modeling.

(4) Generative adversarial learning approaches, such as those discussed in [31, 32, 33], employ a dual-network framework consisting of a generator and a discriminator. The generator learns to model the distribution of normal data, while the discriminator is tasked with distinguishing between normal instances and anomalies. This adversarial training process enhances the system's ability to detect subtle anomalies by leveraging the power of generative models to simulate normal data distributions.

In recent years, some innovative algorithms have emerged in the field of anomaly detection, including: (1) Transformer-based methods [13, 34, 21] harness the architecture of Transformers, which have shown exceptional performance in natural language processing tasks. These models apply self-attention mechanisms that allow them to weigh the importance of different parts of the input data, thus enabling them to capture temporal dependencies and contextual information crucial for identifying anomalies in time series data.

(2) Contrastive learning-based methods [35, 36, 20, 37, 38] focus on learning robust representations by contrasting positive and negative pairs. In the context of anomaly detection, these methods can effectively differentiate between normal and anomalous instances by learning to identify subtle differences in representation, thereby enhancing the model's sensitivity to anomalies. For example, COUTA [35] emphasized calibrating irregular noise and enhancing the significance of normal samples through uncertainty modeling, introducing a one-class loss based on prior distribution to improve prediction reliability.

(3) Diffusion-based approaches [39, 40] utilize mathematical diffusion processes to model how anomalies propagate through complex networks and time series sequences. These techniques are particularly effective in scenarios where the relationships between data points are crucial, as they capture the dynamics of anomaly spread and enhance detection capabilities in interconnected systems. For example, D3R [39] addresses the instability of time series data in non-stationary environments, a common challenge that often leads to a high false positive rate. D3R employs a decomposition and reconstruction method to manage data drift, integrating a noise diffusion model that directly recovers contaminated data. This strategy mitigates the substantial retraining costs typically associated with changes in information bottlenecks, thereby improving overall detection reliability.

(4) Large Language Models (LLMs), such as those adapted for anomaly detection tasks [41], leverage advanced architectures like GPT-2. These models utilize their extensive capabilities in knowledge representation and context understanding to identify anomalies in time series data. By applying techniques from language modeling to time series analysis, LLMs can uncover complex patterns and detect outliers with high accuracy. Generally, the basic structure of these models is consistent with GPT-2. They use self-attention mechanism, feed-forward neural network, and layer normalization to learn sequential features. Their basic units are self-attention module and residual feed-forward neural network, which are stacked multiple times to form the model's decoder. As a pure decoder model, it predicts the next token in an autoregressive way, so an encoder isn't needed. GPT-Adapter [41] just fine-tunes the parameters of the feed-forward neural network and layer normalization. It uses only some of the pre-trained GPT model's parameters to transform text token prediction into time series prediction and assess anomalies based on prediction loss.

B. DETAILED EXPERIMENTAL SETTINGS

We utilize official or open-source baselines that have been published on GitHub. These baselines can be downloaded by following:

- 1) Deep SVDD: <https://github.com/xuhongzuo/DeepOD>
- 2) USAD: <https://github.com/xuhongzuo/DeepOD>
- 3) TCN-ED: <https://github.com/xuhongzuo/DeepOD>
- 4) COUTA: <https://github.com/xuhongzuo/DeepOD>
- 5) TranAD: <https://github.com/xuhongzuo/DeepOD>
- 6) NCAD: <https://github.com/xuhongzuo/DeepOD>
- 7) Deep IF: <https://github.com/xuhongzuo/DeepOD>

- 8) AnomTrans:<https://github.com/thuml/Anomaly-Transformer>
- 9) TimesNet: <https://github.com/xuhongzuo/DeepOD>
- 10) DCdetector: <https://github.com/DAMO-DI-ML/KDD2023-DCdetector/>
- 11) D3R: <https://github.com/ForestsKing/D3R/>
- 12) GPT2-Adapter: https://github.com/PSacf/GPT4TS_Adapter
- 13) NPSR: <https://github.com/andrewlai61616/NPSR/>

A. Datasets details

Existing methods extensively utilize many publicly available real-world datasets. We evaluate model performance on six multivariate datasets:

- 1) MSL (Mars Science Laboratory dataset): Collected by NASA's MSL mission, it reflects various physical phenomena and environmental changes on Mars [57].
- 2) SMAP (Soil Moisture Active Passive dataset): Along with MSL, it's a NASA-collected dataset with telemetry anomaly data from spacecraft monitoring events [58].
- 3) SWaT (Secure Water Treatment Testbed): This platform simulates a small water treatment plant with multiple processing units and sensors, reflecting normal and abnormal states in water treatment [13].
- 4) WADI (Water Distribution Testbed): It simulates a city's water distribution system with pumps, valves, tanks, pipes, and sensors, often used for anomaly detection in industrial control systems.
- 5) PSM (Pooled Server Metrics): From eBay's application server nodes, it has 13 weeks of training and 8 weeks of testing data, covering server performance metrics like CPU and memory usage [59].
- 6) Swan: An IoT-based drinking water quality dataset with a very low anomaly rate, making it a challenging benchmark for anomaly detection.

They can be downloaded in:

- MSL & SMAP: <https://github.com/ML4ITS/mtad-gat-pytorch>
- SWaT: https://itrust.sutd.edu.sg/itrust-labs_datasets
- WADI: https://itrust.sutd.edu.sg/itrust-labs_datasets or https://github.com/Conviss/visual_dataset
- UCR:https://www.cs.ucr.edu/~eamonn/time_series_data_2018/UCR_TimeSeriesAnomalyDatasets2021.zip
- Swan: https://github.com/Conviss/visual_dataset
- Others: <https://github.com/DAMO-DI-ML/KDD2023-DCdetector>.

B. Implementation details

In our default configuration, we set the length of the sliding window to 2048. The patch size is 32. The model is configured with a hidden dimension of 512, 8 attention heads, and 8 layers. We utilize 1000 embedding features. For all methods, we use the optimal F1 search strategy. The training process employs the AdamW optimizer with a batch size 256 and a learning rate of 10^{-3} . We also utilize cosine learning rate scheduling.

The training is conducted over 20 epochs, with each epoch randomly sampling 500 samples. Our algorithm is implemented in Python using PyTorch. All experiments were performed on an NVIDIA A800 (80GB) GPU.

To emphasize the engineering implementation details of the algorithm, we explain the key components in model construction. This section can be directly cross-referenced with the documentation and code comments in our repository. **Positional Embedding:** We use cosine-based positional encoding for temporal positions but do not encode channel position information. The Instance Normalization layer follows the original author's implementation without modifications. **EmbedPatch Encoder:** Each layer contains two transformation matrices, Q and K, implemented using PyTorch's Linear layer with tensor reshaping for multi-head attention. The value matrix V for feature embedding uses PyTorch's native Embedding layer combined with Linear for feature selection. **Reconstruction Layer:** The final layer of SimAD uses a linear transformation to reconstruct features into the original time-series data, recovering all channels of every patch. **ContrastFusion Projection Head:** A two-layer feedforward network (FFN) consisting of Linear \rightarrow ReLU \rightarrow Linear is used. Note that low-dimensional features for each patch must be derived concurrently. **Anomaly Score Construction:** During inference, ensure alignment between the two loss terms. We use linear interpolation to match the length of the cosine similarity score to the MSE score.

C. Complexity Analysis

Assume that batch size is denoted by B , time series length T , channels C , and patch length P , the number of patches is $M = \frac{T}{P}$, the shape of the input can be denoted as $(B, M, P \times C)$. With hidden dimension D , the intermediate feature shape is (B, M, D) . At this point, the computational bottleneck lies in the Transformer's self-attention and FFN. The complexity of self-attention is quadratic concerning the sequence length, with a computational complexity of $(B \cdot M^2 \cdot D)$, while the FFN complexity is $(B \cdot M \cdot D^2)$. The overall computational complexity of the model is $B \cdot (M^2 \cdot D + M \cdot D^2)$. When the number of patches is greater than the model dimension, this complexity can be approximated as $(B \cdot M^2 \cdot D)$. At this point, the length of the time series affects the computational efficiency, leading to slower model performance.

C. ANALYSES OF UAFF AND NAFF METRICS

To demonstrate the reliability of the proposed UAff and NAff metrics both theoretically and practically, we analyze their mathematical properties and practical performance.

A. Mathematical analysis of metrics

1) *Range analysis:* Since NAff is a specific case of UAff, we begin by analyzing UAff. To keep it brevity, we use the following abbreviations: **uaf** for UAff-F1, **uap** for UAff-Pre, **ap** for Aff-Pre, **ar** for Aff-Rec, and **ap_b** for Aff-Pre_{bias}.

Thus, we have $\text{uap} = \frac{\text{ap} - \text{ap}_b}{1 - \text{ap}_b}$. It is theoretically confirmed that both **ap** and **ap_b** fall within the range of [0, 1]. The partial derivatives of **uap** with respect to **ap** and **ap_b** can be given by:

$$\frac{\partial \text{uap}}{\partial \text{ap}} = \frac{1}{1 - \text{ap}_b} \geq 0 \quad (13)$$

$$\frac{\partial \text{uap}}{\partial \text{ap}_b} = \frac{\text{ap} - 1}{(1 - \text{ap}_b)^2} \leq 0 \quad (14)$$

From the above formulas, it can be seen that **uap** increases monotonically with **ap** and decreases monotonically with **ap**_b. When **ap** = 1, **uap** reaches its maximum value of 1. In the case where **ap** = 0 and **ap**_b = 1, **uap** approaches a minimum value of $-\infty$. As concluded in [50], **ap**_b = 1 is applicable only when all the data are anomalies. However, the proportion of anomalies in most datasets is typically below 0.5. Hence, it follows that $\text{ap}_b < \frac{1}{2} + \frac{1}{2} \times 0.5^2 = 0.6125$. Furthermore, the minimum value of **ap**_b tends to approach 0.5. Given that **ap**_b is derived from simulated calculations, we propose a more lenient condition of $\text{ap}_b > 0.4$. In practical scenarios, the range for **ap**_b becomes broader, specifically $\text{ap}_b \in (0.4, 0.6125)$. From Eq.(14), in the case where **ap** = 0 and $\text{ap}_b \rightarrow 0.6125$, **uap** tends towards $\frac{0-0.6125}{1-0.6125} = -1.587$. Therefore, it can be concluded that **uap** falls within the range of $(-1.587, 1]$.

Next, let us proceed with the analysis of **uaf** next. Given that **ar** falls within the range of $[0, 1]$, as indicated by Eq. (11), it becomes evident that when **uap** is within the range of $[0, 1]$, the maximum value of **uaf** is 1. In the case where **ap**_b falls within the interval of $(-1.587, 0)$, it becomes necessary to determine the minimum value of **uaf**. This problem can be reformulated as the search for the corresponding **ap**_b within the range of $(0, 1.587)$ that yield the maximum **uaf**.

The minimum value of **uaf** can be obtained by computing $\frac{-2 \times 1.587 \times 1}{1.587 + 1} = -1.227$. Consequently, the range of **UAff-F1** is $(-1.227, 1]$. Based on the above analysis, it can be concluded that the range of **NAff-F1** is $[-1, 1]$ since it is a specific case of **UAff-F1**.

2) *Distinctiveness of metrics:* It can be briefly analyzed that **UAff** provides greater discrimination than the original **Aff**. The design intention is to ensure that the ratio of **UAff** to **Aff**, $\frac{\text{uap}}{\text{ap}}$, is always greater than 1 or less than 1 in the case of equal **Aff-Pre**. In this case, the ratio depends solely on **Aff-Pre**. Therefore, the problem can be reformulated as proving: $\frac{\text{uap}}{\text{ap}} \geq 1$ or $\frac{\text{uap}}{\text{ap}} \leq 1$.

By calculating the partial derivative of $\mathcal{D} = \frac{\text{uap}}{\text{ap}} = \frac{\text{ap} - \text{ap}_b}{\text{ap}(1 - \text{ap}_b)}$ with respect to **ap**, we obtain:

$$\frac{\partial \mathcal{D}}{\partial \text{ap}} = \frac{\text{ap}_b(1 - \text{ap})}{1 - \text{ap}_b} \geq 0. \quad (15)$$

Hence, when **ap** = 1, $\mathcal{D}_{max} = 1$, indicating that $\mathcal{D} \leq 1$, which aligns with our objective. Based on results in Table III, it is evident that the random algorithm attains the 9th position in **Aff-F1**, whereas it only achieves the 19th and 15th positions in **UAff-F1** and **NAff-F1**, respectively. This observation highlights that the enhanced metrics can provide a more precise evaluation of the algorithm's actual performance. The rankings strongly suggest that the **UAff-F1** and **NAff-F1** metrics offer a superior assessment, compared to the original **Aff-F1**.

B. Bias data of Affiliation precision

Table I presents various datasets' actual and ideal biases. The theoretical ideal bias is determined by assuming the presence of only one anomaly segment in the test data, considering the

anomaly rate and theoretical analysis from [50]. It is noted that the disparity between two biases falls within an acceptable range. In practical scenarios where obtaining the true bias is challenging, the **UAff** can still be computed by estimating the anomaly ratio and calculating the theoretical bias.

$$\text{Bias}_{ideal} = \frac{1}{2} + \frac{1}{2} \cdot (\text{AnomalyRatio})^2 \quad (16)$$

As the occurrence ratio of anomalies is typically low in real-world scenarios, with most anomalies accounting for less than 10% of the data, theoretically, $\text{Bias}_{ideal} < 0.505$. Therefore, in practice, we set $\beta = \text{Aff-Pre}_{Bias} = 0.5$ as the standard parameter for **NAff**. Despite the slight variation from the anomaly rates in publicly available datasets, it is important to note that in the case of real datasets such as Swan, the highest anomaly ratio is 32.6%, resulting in $\text{Bias}_{ideal} = 0.5531$. This scenario is not uncommon, and even in such cases, setting β to 0.5 does not significantly impact the evaluation of different methods. Therefore, setting $\beta = 0.5$ in practice is deemed reasonable. Furthermore, setting $\beta = 0.5$ instead of 0.55 aligns better with real-world scenarios and human cognition, and is more consistent with the original design intent of our metrics.

C. Experimental analysis of **UAff** and **NAff** metrics

Our previous proof compared **UAff** and **NAff** with the original **Aff**, providing empirical and theoretical analyses in Appendix C. We drew inspiration from the ranking-based approach [52]. To highlight the strengths and weaknesses of different metrics, we designed a new experiment.

Experiment Design: We assume a time series of length 1000 with AnomSeq anomaly segments, with a minimum length of MinLen and a maximum length of MaxLen. Using a Random algorithm, we generate uniformly distributed anomaly scores, and then create models with varying performance levels, such as M60, which correctly predicts 60% of normal and anomalous points. Correct anomaly predictions receive a higher anomaly score using a random value U , i.e., $U * 0.1 + 0.9$. Incorrect normal predictions as anomalies also receive higher scores. In other cases, the model outputs an anomaly score of 0. This experiment is repeated 20 times, and the average result is taken.

To investigate different scenarios, we designed three demos in Table VII. Demo1 denotes many anomalies with short intervals. Demo2 denotes a few anomalies with long intervals. Demo3 denotes a few anomalies with extremely long intervals. The score threshold is set to 95%. From the results in this table, we have the following observations:

- 1) **NAff-F1 is better than Aff-F1:** In Demo1 and Demo2, the **Aff-F1** for the Random model are 63.50 and 66.45, which are misleading. Our **NAff-F1** scores are 0.06 and 1.18, indicating that **NAff-F1** is more robust and better at distinguishing random models. Besides, **NAff-F1** has a wider value interval, enhancing its discriminative power.
- 2) **Weak model evaluation:** For the M10 model, which performs worse than a random model, an effective metric should be able to distinguish it from random model. Metrics that achieve this include AUC, **NAff-F1**, **R_A_R**, and **V_ROC**. However, the original **Aff** falls short, assigning a score of 59.38 to M10 in Demo1.

TABLE VII
COMPARISON OF DIFFERENT METRICS ON THREE DIFFERENT SYNTHETIC DATASETS. F1PA IS A POINT-ADJUSTED F1 SCORE.

	Demo	Method	F1	Acc	Pre	Rec	FIPA	AUC	Aff-Pre	Aff-Rec	Aff-F1	NAff-F1	R_A_R	R_A_P	V_ROC	V_PR
Demo1 AnomSeq=5 MinLen=10 MaxLen=12	Random	5.48	90.54	4.60	4.72	26.31	50.25	49.91	87.98	63.50	0.06	77.35	37.85	75.62	36.02	
	M10	2.30	90.05	0.70	0.68	3.45	9.78	46.37	82.99	59.38	-12.08	74.52	36.27	69.24	34.37	
	M60	7.33	90.74	7.30	7.37	39.57	59.75	54.60	88.18	67.18	15.67	74.21	38.71	73.24	36.57	
	M70	10.63	91.00	10.70	10.57	48.15	69.51	52.71	91.76	66.88	9.47	74.08	43.20	74.42	41.50	
	M80	17.90	91.77	18.00	17.83	65.04	80.32	60.68	97.02	74.55	34.03	72.01	43.93	73.45	42.34	
	M90	30.83	93.11	30.70	31.00	72.16	90.04	66.80	98.70	79.56	49.28	70.70	49.95	73.46	49.18	
	M95	50.95	95.12	50.70	51.27	78.37	95.08	75.74	99.50	85.96	67.53	70.10	57.06	73.45	57.52	
Demo2 AnomSeq=1 MinLen=50 MaxLen=60	M100	99.18	99.92	100.00	98.40	100.00	100.00	100.00	99.99	99.99	99.99	69.89	76.48	73.55	79.23	
	Random	4.98	90.42	4.50	4.47	57.40	49.63	50.39	97.82	66.45	1.18	67.02	14.05	65.29	13.84	
	M10	2.29	90.06	0.80	0.73	23.36	10.65	48.99	95.60	64.71	3.73	52.15	10.23	47.37	10.26	
	M60	8.47	90.73	8.10	8.01	65.13	60.45	51.17	98.54	67.28	4.26	66.11	15.77	65.45	15.57	
	M70	11.28	91.11	11.30	11.26	68.92	70.41	54.26	99.25	70.07	14.81	69.77	18.01	69.83	17.79	
	M80	18.66	91.81	18.80	18.53	70.99	79.97	57.17	99.54	72.54	24.20	72.56	23.88	73.50	24.00	
	M90	32.38	93.20	32.60	32.17	74.68	89.97	64.07	99.80	77.91	42.83	74.93	33.12	76.66	33.63	
Demo3 AnomSeq=1 MinLen=300 MaxLen=350	M95	47.65	94.72	48.00	47.34	79.29	94.99	72.65	99.90	84.06	61.87	77.05	45.98	79.01	46.70	
	M100	98.83	99.88	100.00	97.81	100.00	100.00	100.00	100.00	100.00	78.15	80.18	80.47	82.45		

- 3) **Scoring misleadingness:** F1PA is a point-adjusted F1 score that can be misleading when the anomaly interval is long, as seen in Demo3. In such cases, even a random model can achieve a high score of 94.59. Our metrics provide meaningful differentiation, with scores of 19.01, -38.78, and 35.14 for Random, M10, and M60, respectively, distinguishing between these models clearly.
- 4) **Parameter selection:** R_A_R requires selecting an interval length, which can lead to inaccurate evaluations when anomalies are short, as demonstrated in Demos 1 and 2. In Demo 1, the R_A_R scores for Random, M10, and M60 are all above 74, showing that the metric fails to distinguish between these models. This issue is mitigated with longer anomaly intervals, as seen in Demo 3, where M10 has a R_A_R score of 25.81, significantly lower than 50. In contrast, our metric does not require parameter selection and provides accurate evaluations across various scenarios.
- 5) **Weakness of point-based evaluation:** F1, Acc, Pre, and Rec are point-based metrics that are easily influenced by threshold selection and strict label matching, resulting in suboptimal results, even for perfect models like M100. In contrast, event-based metrics are not affected.

Based on these studies, we demonstrate the advantages of our metric over previous ones. We recommend using NAff-F1 and VUS_PR in the future, and other VUS series metrics can also be used when the anomaly interval is long.

D. INVESTIGATION OF EXISTING EVALUATION METRICS

A. Flaws in previous metrics

Fig. 9 illustrates an artificial example of time series anomaly detection. The top row represents the true labels of the time series data. Random1, Random2, Pred1, and Pred2 denote the anomaly scores generated by two random and other algorithms. The labels ending with “PA” indicate point adjustment, where if the model predicts any part of an anomaly segment, the entire segment is considered an identified anomaly.

Random1 predicts all timestamps as anomalies, while Random2 uniformly and randomly predicts anomalies. However, both algorithms achieve high point-adjustment F1 scores (F1PA), indicating that F1PA lacks robustness in this case.

The F1 score can exhibit this issue, as it is more suitable for evaluating shorter interval anomalies (such as the last). However, its results can still be informative and provide some reference when facing non-random predictions.

Pred1 is the top-performing model among the mentioned ones. However, its Aff-F1 score is lower than that of Random1. This is because the original Aff-F1 metric does not account for false predictions. However, the enhanced UAFF-F1 metric provides a better basis for comparing the performance of Pred1 and Random1. Additionally, Pred1 outperforms Pred2, although the difference in their Aff-F1 scores is small and even lower than that of Random1. In this case, the UAFF-F1 metric can provide a more effective comparison between the two models, yielding scores of 0.1818 and 0.0057 for Pred1 and Pred2, respectively, while assigning lower scores to Random1 and Random2.

Based on the above analysis, we can conclude that F1 can serve as a valuable reference metric for non-random algorithms. F1PA may result in false improvements, whereas Aff-F1 lacks sufficient discriminative power and provides inaccurate evaluations for random algorithms. Ultimately, we recommend using the enhanced UAFF-F1 metric for a more accurate performance assessment.

B. Limitations & Advantages of metrics

We referenced the study VUS [52] and analyzed the attributes of various metrics, focusing on discrimination and semantics as key evaluation criteria. It is important to note the following properties of existing evaluation metrics:

- 1) Non-Threshold: This indicates that the metric does not require the setting of a threshold.
- 2) Sequence: This signifies that the metric can effectively evaluate sequential anomalies.

Label	0	0	0	0	0	1	1	1	1	0	0	0	0	1	P	R	F1	F1PA	Aff-F1	NAff-F1			
Random1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.375	1.0	0.546	0.546	0.7218	0.0635			
Random2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.3750	0.500	0.429	0.625	0.6623	-0.0613			
Random2PA	1	0	1	0	1	0	1	1	1	1	1	0	1	0	1	0	0.3333	0.167	0.222	0.769	0.6882	0.1918	
Pred1	0.1	0.3	0.2	0.1	0.3	0.7	0.2	0.8	0.2	0.3	0.4	0.2	0.2	0.8	0.1	0.3333	0.167	0.222	0.769	0.6882	0.1918		
Pred1PA	0	0	0	0	1	0	1	1	1	1	1	1	0	0	0	1	0	0.3333	0.167	0.222	0.769	0.6882	0.1918
Pred2	0.3	0.1	0.2	0.1	0.1	0.6	0.2	0.1	0.7	0.1	0.4	0.1	0.6	0.2	0.8	0.2	0.2500	0.167	0.200	0.714	0.6622	0.0057	
Pred2PA	0	0	0	0	0	1	1	1	1	1	1	1	0	1	1	1	0	0.2500	0.167	0.200	0.714	0.6622	0.0057

Fig. 9. A case of artificial data illustrating the shortcomings of different metrics.

TABLE VIII
THE PROPERTIES OF DIFFERENT METRICS.

Metric/Property	Score Threshold	Sequence-adapted	Parameter-free	Discrimination	Semantics
ACC	x	x	x	x	x
F1	x	x	x	x	x
FIPA	x	✓	x	x	x
AUC	✓	✓	x	✓	✓
Aff-F1	✓	✓	✓	x	✓
UAff-F1	✓	✓	✓	✓	✓
R_A_R	✓	✓	x	✓	✓
R_A_P	✓	✓	x	✓	✓
V_ROC	✓	✓	✓	✓	✓
V_PR	✓	✓	✓	✓	✓

- 3) Parameter-Free: This means the metric does not require additional parameters for its evaluation.
- 4) Discrimination: This attribute reflects the metric's ability to differentiate between random, weak, and strong models. For example, our research found that the F1PA metric exhibits low discrimination, as even a random model could achieve a score of 94.
- 5) Semantics: This denotes that the metric is associated with human-understandable meaning. For instance, the AUC metric provides clear semantics for point anomalies or short anomalies. Similarly, the Aff metric offers semantic relatedness by indicating how closely the detected anomalies align with the actual anomaly range—a higher score suggests a closer match.

These attributes are critical for evaluating the effectiveness of different metrics in anomaly detection. Table VIII summarizes the properties of the different metrics.

1. Affs (including the metrics we proposed) are based on local affiliation, which enhances their ability to evaluate a model's early warning and post-alert performance. 2. The VUS series of metrics are designed based on AUC, with a focus on evaluating a model's tendency towards anomalies and normal events. In contrast, our metrics and VUS represent two distinct approaches to evaluation. 3. For a comprehensive comparison of other metrics, please see Appendix C-C. Below, we summarize the key conclusions:

- 1) NAff-F1 has a better discriminative ability than Aff-F1, which was our initial design goal.
- 2) ACC, F1PA, and Aff exhibit insufficient weak model evaluation capabilities and cannot effectively evaluate random models and weak models.
- 3) Scoring misleadingness: F1PA has the strongest misleading score.
- 4) Refer to the table above; some metrics require additional parameters. As Paparrizos et al. [52] suggests: “We observe that this change implies a larger variation for several

threshold-based measures. Thus, the latter confirms the limitations and the non-robustness of threshold-based measures to the anomaly cardinality ratio.” Threshold-based metrics have certain limitations.

- 5) Weakness of point-based evaluation: Previous research [60] has shown that current datasets may have unreasonable labels. When using point-based metrics, the weaknesses of the evaluation metrics are magnified. Strict label matching can lead to inaccurate evaluations.

These observations highlight the complexity of real-world scenarios and the varied emphasis of different metrics. In the TSAD community, interval evaluation is currently deemed more crucial, which is why we prioritize Aff and VUS metrics. As the field of TSAD is diverse, no single evaluation metric is universally appropriate, and the choice of metric should be made carefully for each case [61].

In summary, it is currently recommended that “when publishing results in TSAD research, multiple metrics should be included, and both the code and the anomaly scores should be made available to facilitate easy comparison with any evaluation metric.” [61] A more comprehensive and systematic investigation would require a broader research scope, which is beyond the focus of this paper. To contribute to the advancement of TSAD, we have made our model code, weights, and training and testing scripts available in order to reproduce results reported in our work.

E. DETAILED RESULTS OF EXPERIMENTS

Table IX presents the performance of three models, namely SimAD, AnomTrans, and DCdetector, on the UCR dataset. Considering the overall performance across all sub-datasets, it is evident that the SimAD model consistently achieves higher F1 scores compared to the baseline models. Additionally, the SimAD model for anomaly detection demonstrates the highest number of datasets with positive evaluation scores. By excluding datasets with negative evaluation scores, SimAD exhibits significantly higher composite scores for all evaluation metrics than the other baseline models. These findings indicate that SimAD can detect a wider range of anomalies. The specific results of the ablation experiments are presented in Table V.

TABLE IX
COMPARISON RESULTS ON UCR DATASETS.

Metrics	Methods	Datasets					
		F1	Aff-Pre	Aff-Rec	Aff-F1	UAff-F1	NAff-F1
Count	AnomTrans	147	244	244	244	141	148
	DCdetector	149	206	206	206	127	144
	Ours	240	240	240	240	173	177
Avg.+	AnomTrans	1.17	50.66	98.94	66.88	9.41	10.88
	DCdetector	1.66	50.90	99.96	67.41	6.71	5.52
	Ours	14.99	57.83	99.91	72.52	35.16	34.38
Avg.	AnomTrans	1.15	50.39	97.73	66.55	1.30	1.23
	DCdetector	1.61	50.63	83.30	67.06	2.01	2.64
	Ours	14.99	57.83	97.47	72.52	19.53	1.00

A. Additional results of comparison

Since the VUS series evaluates anomaly detection models differently, we included four VUS metrics in the comparative

TABLE X
PERFORMANCE OF DIFFERENT METHODS ON VUS METRICS.

Datasets	MSL				SMAP				SWaT				WADI				PSM				NIPS-TS-Swan				
Methods	R_A_R	R_A_P	V_ROC	V_PR	R_A_R	R_A_P	V_ROC	V_PR																	
Random	57.27	14.15	57.56	14.66	57.59	17.20	52.71	14.14	57.29	16.26	51.21	12.74	56.81	7.70	58.35	8.58	55.72	33.01	55.09	32.59	38.17	35.43	86.19	83.58	
LOF	62.15	17.29	60.85	17.19	58.00	18.63	57.56	18.59	66.94	42.33	66.67	41.88	46.17	7.88	45.23	7.73	81.36	65.24	80.42	64.55	90.34	89.83	89.42	88.56	
IForest	65.38	18.37	64.74	18.29	59.30	16.46	59.31	16.45	73.25	53.11	72.29	52.07	78.99	28.47	77.89	27.02	75.81	55.92	75.19	55.50	91.88	92.00	91.27	90.98	
PCA	59.81	18.20	59.42	18.08	44.03	11.69	43.90	11.69	61.61	44.38	61.87	44.75	48.85	7.83	48.25	7.72	74.52	54.93	73.97	54.55	91.50	91.64	90.42	90.23	
Deep SVDD	60.15	19.44	59.54	19.36	44.03	11.55	41.95	11.56	50.96	56.48	56.48	57.21	54.82	57.16	54.73	54.00	66.98	53.97	66.92	82.27	92.73	80.89	91.74		
USAD	59.59	20.00	58.86	19.77	33.76	11.27	33.78	11.28	61.41	63.41	84.38	62.74	50.82	54.31	50.78	54.24	64.47	69.55	64.02	69.18	93.70	96.14	91.40	94.50	
TCN-ED	58.98	18.20	58.62	18.13	43.93	11.73	43.79	11.73	61.53	57.95	61.53	57.94	50.86	54.31	50.83	54.24	58.20	43.02	57.48	42.64	82.27	92.72	80.88	91.73	
COUTA	61.97	18.91	61.49	18.68	44.82	11.91	44.72	11.91	72.68	25.09	72.57	25.06	46.47	30.27	45.65	29.43	64.73	43.29	64.50	43.15	83.78	93.36	82.15	92.26	
TranAD	57.77	17.67	57.25	17.55	45.27	11.98	45.15	11.98	86.59	64.29	86.25	64.19	60.27	44.03	59.94	43.65	72.85	56.28	72.42	56.02	89.09	93.61	88.01	92.59	
NCAD	67.98	21.47	67.29	21.32	48.39	13.21	48.37	13.22	76.86	44.96	76.76	44.91	45.80	8.51	44.52	8.13	67.60	48.55	67.09	48.26	87.40	94.73	85.31	93.44	
Deep IF	53.19	17.24	52.33	17.04	56.93	14.77	56.89	14.77	50.19	56.40	50.19	56.39	47.97	13.58	47.07	12.23	68.37	46.48	67.71	45.98	82.27	92.73	80.89	91.74	
AnomTrans	53.25	13.17	53.03	13.21	52.75	16.56	52.61	16.55	21.52	8.29	21.49	8.29	50.49	7.83	50.29	7.71	49.84	31.66	49.12	31.56	85.46	83.57	83.56	81.64	
TimesNet	65.41	20.95	64.63	20.67	49.01	12.93	48.83	12.91	31.27	10.31	31.10	10.29	81.75	34.76	79.80	32.62	67.22	49.17	66.37	48.05	92.61	95.53	91.58	94.30	
DCDetector	52.17	14.69	52.00	14.68	48.65	16.88	48.68	16.88	50.85	14.70	50.85	14.70	44.68	51.62	8.24	51.60	52.56	32.56	51.73	32.56	87.59	85.84	85.84	84.31	
D3Net	69.74	21.04	69.02	20.93	54.98	16.56	48.67	16.56	71.03	39.91	84.38	37.27	66.43	7.49	49.25	5.87	64.47	40.44	66.92	49.38	52.46	41.59	89.47	89.27	88.29
GPTA-Adapter	52.97	17.79	53.57	17.38	59.90	19.03	57.31	17.58	51.11	14.10	50.01	10.29	49.87	5.82	51.52	4.82	48.24	36.56	32.79	47.37	41.59	84.31	84.23		
NPSR	67.58	19.18	66.44	19.72	47.69	14.22	41.91	11.42	87.52	67.26	84.93	65.42	87.94	54.45	88.23	56.72	72.57	58.65	70.55	52.12	53.92	47.61	90.90	90.35	
M2N2	65.82	21.69	65.40	21.55	48.80	14.36	48.74	14.36	14.36	61.41	26.33	61.14	26.16	45.07	8.15	43.99	7.98	39.19	30.22	39.20	30.10	63.95	29.51	63.69	29.15
AdaMemBLS	58.29	16.13	57.46	16.06	55.73	14.26	55.57	14.26	84.27	68.64	82.00	66.12	66.50	23.11	63.47	21.40	66.16	49.81	65.07	49.24	88.85	87.85	87.12	85.92	
Ours	69.45	24.86	68.45	24.15	61.90	20.77	58.64	17.02	86.19	69.46	83.89	66.70	91.26	61.99	91.95	64.32	71.25	54.79	68.06	53.81	87.59	85.09	90.95	91.74	

experiments: Range-AUC-ROC (R_A_R), Range-AUC-PR (R_A_P), V_ROC, and V_PR. The metrics V_ROC and V_PR assess performance based on the surfaces created by ROC and PR curves. As shown in Table X our method achieves the best performance on MSL, SMAP, SWaT, and WADI, and demonstrates strong competitiveness on PSM and Swan. This can be attributed to the anomaly ratios of dataset and evaluation metrics. Specifically, PSM and Swan have a higher ratio of anomalies, which may impact performance. Overall, SimAD is the most stable, performing consistently well across multiple metrics and datasets, compared with other methods.

B. Additional visualization analysis

Fig. 10 demonstrates the impact of different loss functions on the predicted results, threshold values, and anomaly scores. The figure shows that our model's performance is enhanced when both the L2 and cosine loss functions are employed, as compared to using either the L2 loss function alone or the cosine loss function alone.

Fig. 11 illustrates the actual detection performance of the AnomTrans and NPSR on the SWaT dataset. AnomTrans can only detect a few anomalies. In Case 1, two anomalies were not detected and AnomTrans struggled with accurately detecting segment anomalies. Moreover, the anomaly score differentiation is relatively low, as in Case 2, where it has limited success detecting anomalies. In cases 3 and 4, NPSR is more effective in detecting anomalies than AnomTrans, especially segment anomalies. However, NPSR fails to detect certain anomalies, such as near time points 6500 and 8000. Additionally, NPSR tends to produce more false positives, as observed around time point 140000, where a few instances are falsely classified as positive. Furthermore, Fig. 12 illustrates the similarity scores at various model layers. In the initial layer, the similarity scores exhibit a relatively dispersed pattern. However, as the number of layers increases, the similarity scores gradually rise, leading to more focused attention.

C. Analysis of cosine similarity loss

1. Direct Explanation: MSE enforces point-wise reconstruction by requiring the model to return each time point in the sequence to its original value, acting as a point-wise loss. Cosine similarity, on the other hand, enhances the alignment between

patches, promoting smoother transitions and reducing noise interference during reconstruction. **2. Semantic Explanation:** Patch-based time series studies suggest that patches are more effective at capturing the semantic information of sequences. While interpolative learning can reconstruct individual time points, patches benefit from more than just simple interpolation. Traditional patch learning often relies solely on MSE loss, ignoring this benefit. **3. Visual Explanation:** As shown in Fig. 10, incorporating cosine loss enables the model to better detect longer anomalies. For additional details, see Section E-B. **4. Numerical Explanation:** Ablation experiments were conducted by removing cosine terms from Equations 4, 5, and 7 while keeping the MSE loss. The results, presented in Table V, indicate a significant decrease in model performance after removing these terms.

D. Detailed quantified comparison

For a detailed comparison, we group Deep SVDD, Deep IF, USAD, TCN-ED, NPSR, and TimesNet, as they can be considered traditional deep learning methods. The remaining methods are grouped as in Section IV-A. Below is the analysis and discussion.

1. In the traditional deep learning group, NPSR has achieved the highest average ranking. Our proposed method significantly outperforms NPSR on datasets SWaT, WADI, and Swan. All of these dataset are more complex and higher-dimensional. The superiority of our proposed method over NPSR can be attributed to that the former can handle high-dimensional data better and have a larger parameter capacity.

2. Among Transformer-based models, TranAD has obtained the highest average ranking. With more parameters, TranAD performs better on datasets SWaT, PSM, and Swan. However, since these models rely on point-wise modeling of time series data, their complexity is high, limiting their performance. In contrast, SimAD employs a patch mechanism, allowing the model to support window sizes of 1024 or higher, resulting in a larger receptive field and the ability to capture more information. **SimAD's patch mechanism enables larger window sizes and better information capture compared to point-wise Transformer models like TranAD.**

3. Contrastive learning-based methods perform similarly on average. Taking COUTA as an example, it utilizes multiple methods to generate negative samples and leverages contrastive

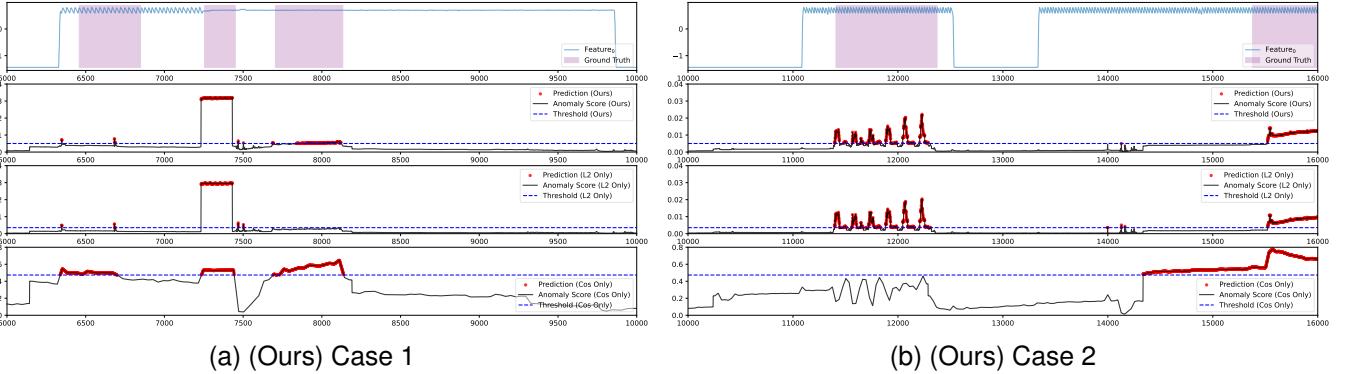


Fig. 10. SimAD's detection performances in real-world.

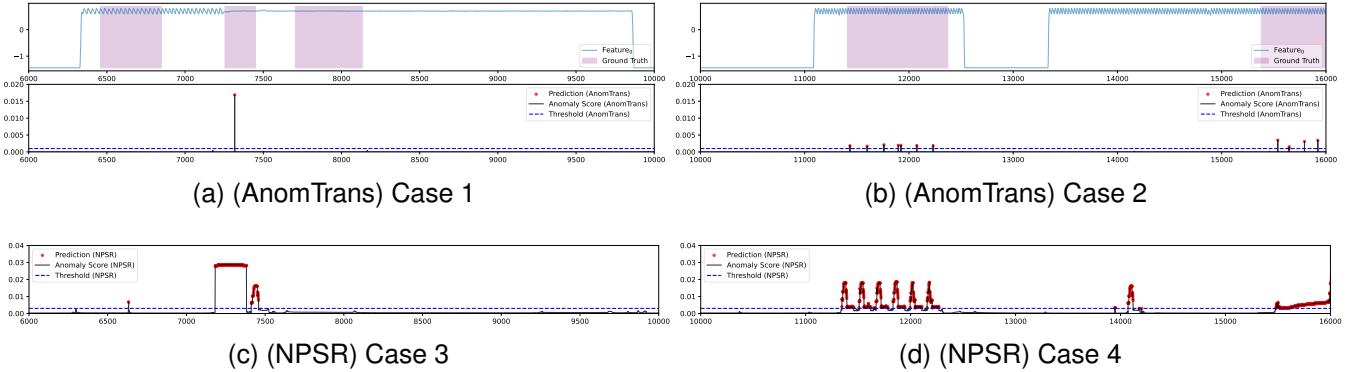


Fig. 11. Other algorithms' detection performances in real-world.

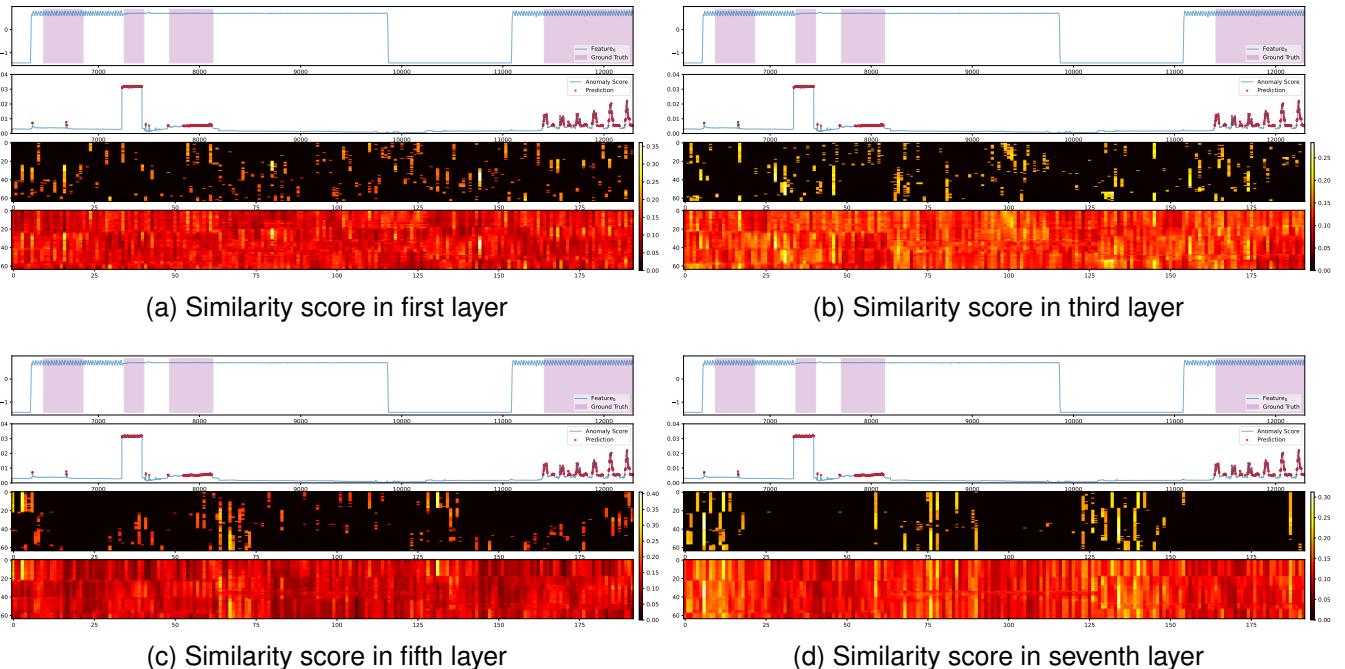


Fig. 12. Similarity scores in different layers.

learning to enhance model robustness. However, its reliance on a TCN network structure and complex sample generation strategy necessitates high-quality datasets. This results in superior performance on MSL and SMAP but causes sharp declines on other datasets. In contrast, our method employs a simpler negative sample generation strategy and a more complex network structure, making it less dependent on dataset quality. **Our method's simpler negative sample generation and more complex network structure make it more adaptable to various dataset qualities compared to contrastive learning methods like COUTA.**

4. D3R is the only **diffusion model**. It preserves stable components in time series data, such as trends and periodic information, which enhances anomaly detection, particularly for unstable time series data with sudden increases. However, its performance is limited by stronger prior assumptions and data quality requirements, making it less effective across all datasets. In contrast, SimAD operates on more general assumptions, recognizing that normal and anomalous data have different representations. This universal modeling approach allows SimAD to adapt more effectively to a wide range of datasets. **SimAD's general assumptions about data representation enable it to adapt more effectively across various datasets compared to the diffusion model D3R.**

5. GPT2-Adapter is an **LLM-enhanced** time series anomaly detection model that integrates large language models with time series data. Although it represents a pioneering effort in applying LLMs to this field, its performance in anomaly detection falls short. This is mainly due to challenges with aligning natural language models with time series data, difficulties in fine-tuning LLM parameters, and the lack of specialized optimization for time series anomaly detection. Consequently, GPT2-Adapter often underperforms in practical scenarios. **Despite being a pioneer in LLM-based approaches, GPT2-Adapter falls short due to modal alignment issues and insufficient optimization for time series anomaly detection.**

F. TIME COSTS

Table XI presents a comparative analysis of SimAD and other models regarding model size, training time, and testing time using the SWaT dataset. The estimated training time for all methods involved training on 128,000 samples to ensure a fair comparison. However, an issue with the original code of the D3R model caused its reported testing time to be excessively long (2794.25s). To provide a more accurate estimation, we estimated the actual testing time of D3R to be 15.98s.

In terms of model size, SimAD demonstrates a significant advantage over D3R and GP2-Adapter (LLM model). Its smaller model size indicates efficient memory usage. SimAD also performs well regarding inference time, showing shorter inference times than other models. This implies that SimAD can make predictions faster during deployment, which can be crucial in real-time applications.

However, it's important to note that SimAD has a longer training time and a larger model size than the other models in the analysis. Despite this, considering that SimAD achieves performance improvements of over 60%, the model size and

inference speed of SimAD are still within an acceptable range. To further improve SimAD, future work will explore model optimization techniques to reduce the model size while maintaining or enhancing its performance. This would address the longer training time and larger model size, making SimAD even more efficient and practical for deployment.

TABLE XI
THE TIME COSTS OF DIFFERENT ALGORITHMS. THE LOWER IS BETTER.

Methods	Model Size (MB)	Train (sec)	Test (sec)
USAD	332.07	72.14	1.51
TCN-ED	0.04	41.34	0.86
NCAD	0.19	141.73	1.77
AnomTrans	28.29	33.25	3.91
TimesNet	19.15	24.29	3.62
D3R	225.10	1900.37	15.98
GPT2-Adapter	241.58	241.58	71.24
Ours	114.75	245.88	1.27

G. BORDER IMPACTS

Time series anomaly detection is crucial in identifying abnormal behaviours and events. Its applications span across various domains, including risk prevention, health monitoring, and industrial inspection. The demand for effective anomaly detection techniques is high due to its wide industrial coverage. In practical scenarios, labeled data for anomalous events is often scarce. However, our method is designed to adapt to this situation and provide highly accurate detection results. There are scenarios where the model needs to perform well under resource-constrained conditions. Therefore, it becomes necessary for our model to incorporate other techniques that can reduce its size and enable it to adapt to such situations. In the future, we aspire to address this limitation by considering the model lightweight while ensuring detection accuracy, thereby creating greater societal value.

H. LIMITATIONS

Significant computational costs are incurred due to the adoption of a Transformer-like architecture as the backbone of the model. In the future, we will explore alternative network frameworks, such as MAMBA or convolutional networks, that are more lightweight and efficient. Furthermore, considering the existing limitations of current time series anomaly detection algorithms, such as inaccurate evaluation and lengthy computation, we aim to propose novel evaluation metrics that are both accurate and efficient for anomaly detection.

Below is the in-depth analysis of the intrinsic shortcomings of the model.

TSAD in Practical Scenarios: TSAD often requires fast and effective models. Our Transformer-based model's computational cost is a limitation. Assuming batch size B , time series length T , channels C , and patch length P , the number of patches is $M = \frac{T}{P}$, and the final input is $(B, M, P \times C)$. With hidden dimension D , the intermediate feature shape is (B, M, D) . At this point, the computational bottleneck lies in the Transformer's self-attention and feed-forward network (FFN) [45]. The complexity of self-attention is

quadratic concerning the sequence length, with a computational complexity of $(B \cdot M^2 \cdot D)$, while the FFN complexity is $(B \cdot M \cdot D^2)$. The overall computational complexity of the model is $B \cdot (M^2 \cdot D + M \cdot D^2)$. When the number of patches is greater than the model dimension, this complexity can be approximated as $(B \cdot M^2 \cdot D)$. Note that SimAD’s ContrastFusion is a two-layer FFN projection head, whose complexity equals that of the FFN. During reconstruction, an extra linear layer with lower complexity than the FFN is used, and their complexities differ by a constant factor. Thus, SimAD’s overall complexity is dominated by the EmbedPatch Encoder’s complexity, with other components contributing less. At this point, the length of the time series affects the computational efficiency, leading to slower model performance.

To address the computational cost, we can replace the attention backbone in SimAD with frameworks like SSM or MAMBA, whose complexity is linear concerning the sequence length, resulting in $(B \cdot M \cdot D^2)$ overall, with FFN as the bottleneck. However, when N is less than the model dimension, linear complexity models do not resolve the bottleneck. In this case, we can explore other FFN structures, such as using grouped Linear layers, dividing the input into G groups of size $\frac{D}{G}$. This reduces the FFN complexity to $(B \cdot M \cdot \frac{D^2}{G})$.

Comparison with Non-Patch-Based Transformer Models:

For non-patch-based Transformer models like AnomTrans, the computational bottleneck lies in the self-attention calculation, leading to a complexity of $(B \cdot T^2 \cdot D)$. Since we slice the sequential data, the actual sequence length in our network is shorter. Therefore, our computational complexity is more efficient.

Comparison with Channel-Independent and Patch-Based Models: For channel-independent, patch-based models like DCdetector, the complexity is $(B \cdot C \cdot M^2 \cdot D)$, where B is the batch size, C is the number of channels, M is the length of the patch sequence, and D is the dimensionality of features. These models must process each channel independently, resulting in higher computational complexity. Consequently, our approach offers a more efficient alternative in this respect.

I. INSIGHTS BEHIND SIMAD

1. Our model emphasizes simplicity & effectiveness, avoiding excessive complexity. Our goal is to establish a new paradigm for time series anomaly detection (TSAD), akin to how SimCLR sets the foundation for image representation. We consider that a simpler, more effective model allows for further exploration and development, leaving ample room for future improvements to SimAD. For example, SimAD employs the simplest noise augmentation for generating negative samples and utilizes a straightforward MSE loss for constructing an asymmetric optimization strategy, yet it still delivers excellent results. We expect that SimAD’s focus on simplicity and core principles will make it a cornerstone of future TSAD research, offering new perspectives to the community.

2. Our proposed UAFF and NAAFF are simple & effective metrics. Many related works have highlighted the shortcomings of existing TSAD metrics. While the Affiliation metric provides a solid foundation, it has notable deficiencies: it fails

to accurately evaluate random algorithms and weaker methods, and its discrimination is limited, overlooking performance differences between models. We believe that simplicity and effectiveness are central to our approach, as reflected in our improvements to the metrics.

3. We fully leverage the concept of viewing similarity from different perspectives, operating under the belief that normal and abnormal time series representations are inherently dissimilar. Although SimAD may superficially resemble BYOL [62] in terms of its loss function design, the underlying motivations differ significantly. BYOL is based on the premise that representations of different views of the same data should be similar in the shared latent space. This is achieved in BYOL by employing contrastive learning to bring representations of different views closer together. Besides, BYOL updates its parameters via exponential moving averages. In contrast, SimAD employs a strategy that emphasizes maximizing the dissimilarity between representations of normal data and those of abnormal data. Furthermore, our model updates its parameters through an asymmetric loss optimization objective function, guided by a stop-gradient mechanism. SimSiam [49] is a self-supervised learning framework that enhances image generalization by learning view-invariant representations, using noise as a form of real-world augmentation. It utilizes two networks without parameter sharing and employs a cosine similarity loss to align projected and pre-projected features, effectively preventing collapse to trivial solutions. In contrast, our proposed method, SimAD, is designed to distinguish between normal and anomalous data by encouraging the projected features of two views to diverge in a low-dimensional space. This key difference in both implementation and objective distinguishes SimAD from SimSiam.

4. The motivation of negative samples: Previous research (Sec. 4.5.1 in [9]) has shown that adding noise and scaling to time series data can alter the underlying information within the series. Study [35] employed a single time point alteration strategy to generate synthetic anomalies (contaminated data), while [63] used contrastive learning-based single-class algorithms. Both approaches indicate that adding noise can create varying perspectives, thereby helping to clarify the model’s decision boundaries.

The distinction between normal and anomalous instances is influenced by previous research, which indicates that adding noise can distort the original information in time series, potentially generating synthetic anomalies. This is intuitive, as increased noise leads to greater deviations from the original patterns. While TSAD currently does not use labels during training, research has shown that contrastive learning with negative samples can enhance a model’s decision boundaries. As the gap between original time series data and augmented data widens, the model becomes better at capturing consistent information, such as trends and periodic patterns.

Our experiments and visualizations suggest that EmbedPatch primarily learns from normal data, resulting in less resemblance to anomalous time series. It is acknowledged that simple noise may not capture all the complexities of real-world anomalies. Future research aims to develop strategies that better address various real-world anomalies, as exemplified

in [63]. Nonetheless, our paper presents a more foundational approach using simpler strategies, offering a fresh perspective for the TSAD community.

In time series anomaly detection, [37] introduces noise into the embedding and tasks the model with denoising to reconstruct the original time series. Similarly, in image anomaly detection, [11] demonstrates the effectiveness of “adding perturbations to feature tokens, guiding the model to learn knowledge of normal samples through the denoising task.” Building on this prior work, we assert that the primary function of denoising is to guide the model in learning the patterns of normal data, implicitly directing its focus towards normal behavior. Our ablation experiments further validate this point, showing that the model’s performance deteriorates when denoising is not applied.

5. The concept of “Dissimilarity” refers to the dissimilarity between normal and abnormal samples or representations. Specifically, this concept manifests in the following aspects:

- **Dissimilarity between positive and negative samples:**

We expand the representation space between positive (original time series) and negative (constructed using noise) samples, thereby reducing their similarity and increasing their dissimilarity.

- **Dissimilarity between normal and abnormal samples:**

By diminishing the similarity between positive and negative samples, our objective is to heighten the differentiation between normal and abnormal samples, thus enabling the model to enhance anomaly detection efficacy. The distinction between normal and abnormal sample representations forms the core motivation behind SimAD. As no labels are accessible during training, it is hard to determine whether a sample is anomalous. Nonetheless, prior research as well as our own work indicate that even in the absence of labels, SimAD can effectively model the dissimilarity between normal and abnormal instances through learning the dissimilarity of positive and negative samples.

- **Dissimilarity embedded in EmbedPatch:**

EmbedPatch, a crucial component in our model, introduces learnable parameters independent from input data. These parameters enable the model to capture the most prevalent features in data and increase the difficulty of reconstruction, compelling the model to learn the universal characteristics of time series. A higher-layer EmbedPatch acquires more abstract normal semantic information, while a lower-layer EmbedPatch extracts simpler patterns from the time series.

- **Contrastive learning design utilizing dissimilarity:**

We consider normal and abnormal time series as having dissimilar view representations. Equation 7 uses MSE and cosine similarity to separate the representation spaces of normal and abnormal data, thereby enlarging their dissimilarity. While our approach based on dissimilarity shares some similarities with the concept of contrastive learning to a certain extent, there are fundamental differences in principles and implementation goals. From a theoretical perspective, contrastive learning often assumes that the augmented samples are similar to the original samples,

with other dissimilar samples serving as negative samples. However, based on the findings of previous studies [63], considering augmented samples as negative samples in time series, i.e., dissimilar to the original samples, aligns better with the characteristics of time series data. Therefore, in principle, our dissimilarity-based approach differs from contrastive learning methods. In terms of implementation, our dissimilarity-based approach places greater emphasis on the dissimilarity between patches and the dissimilarity between normal and abnormal instances, utilizing cosine similarity to characterize this relationship. This implementation aspect also sets our approach apart from contrastive learning methods, hence the designation of our method as the “dissimilarity-based” approach seems more appropriate.

J. DISCUSSION

How does patch-based feature extraction work? Patch-based feature extraction methods are not new in the context of time series, and they have been effectively verified in time series prediction and anomaly detection in the past. For example, in PatchTST [55], the authors use the patch approach to segment the time series, and then use an attention model to learn the time series features and predict future time series data. In DCdetector [20], the authors also adopt the patch-based method to model the local and global time series information. In this way, the model is guided to capture the differences between normal and abnormal time series in both global and local aspects, and finally, these features are utilized to complete the anomaly detection. However, in previous papers, there is a lack of in-depth exploration of the patch-based approach. Mainly, these methods are limited by the size of the time window. Usually, the time window size they set by default is 105, and the size of the patch is often 3, with a maximum of 7. Under such settings, the intuitive understanding is that they use three time points to represent the features of a very small interval of the time series. In fact, this method does not have a significant difference from the past modeling methods that are completely based on the time window. Because ultimately, the model is still limited by the time window of 105, which makes the model only learn the time series features in a local area. But in the real world, time series data often contains at least tens of thousands or even hundreds of thousands of timestamps. Therefore, a time window of 105 makes it difficult for the model to learn longer context relationships. To address this issue, for the first time, we scale the time window to 2048, which is 20 times larger compared to the previous models. However, this expansion comes at a cost. Since a longer time window requires the model to process more data at once, it will lead to extremely slow model inference speed, which is at least one-twentieth of the previous speed. And in the actual inference scenario, this is completely unusable. To solve this problem, we make improvements in the way of time series feature extraction, that is, we change the fixed mindset of using the original patch. We also expand the size of the patch to at least 32 lengths. That is to say, we use the data of 32 time points for representation. Finally, the model only needs to process 2048/32 basic units,

so the features actually processed by the model are reduced. At the same time, due to the existence of a lot of redundant data in the time series itself, this processing method does not limit the model's learning of time series data. For example, in some time series, the channel values of the time series may remain around a certain level for a long time, with only slight fluctuations. In this case, the truly effective information is to mine the stable values of the time series during this period. Or, in the time series, periodic information is often very common, and there may be slight changes within each period. Similarly, most of the time series data is redundant at this time. Therefore, the patch-based processing and expanding the size of the patch to 32 will not affect the model's learning of time series data. Meanwhile, our experiments also show that when the time window is larger and the patch is larger, the loss of the model is actually smaller, which means that the model fits the time series more perfectly.

REFERENCES

- [1] Z. Zhong, Z. Yu, Z. Fan, C. L. Philip Chen, and K. Yang, "Adaptive memory broad learning system for unsupervised time series anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 5, pp. 8331–8345, 2025.
- [2] Z. He, P. Chen, X. Li, Y. Wang, G. Yu, C. Chen, X. Li, and Z. Zheng, "A spatiotemporal deep learning approach for unsupervised anomaly detection in cloud systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 4, pp. 1705–1719, 2023.
- [3] Y. Zheng, H. Y. Koh, M. Jin, L. Chi, K. T. Phan, S. Pan, Y.-P. P. Chen, and W. Xiang, "Correlation-aware spatial-temporal graph learning for multivariate time-series anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 9, pp. 11802–11816, 2024.
- [4] Z. Yu, Z. Zhong, K. Yang, W. Cao, and C. P. Chen, "Broad learning autoencoder with graph structure for data clustering," *IEEE Transactions on Knowledge and Data Engineering*, 2023.
- [5] B. Kim, M. A. Alawami, E. Kim, S. Oh, J. Park, and H. Kim, "A Comparative Study of Time Series Anomaly Detection Models for Industrial Control Systems," *Sensors*, vol. 23, no. 3, p. 1310, 2023.
- [6] K. Zhu, P. Song, and C. Zhao, "Fuzzy state-driven cross-time spatial dependence learning for multivariate time-series anomaly detection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2024.
- [7] Z. Yang, R. Zhao, X. Meng, G. Yang, W. Sun, S. Zhang, and J. Li, "A multi-scale mask convolution-based blind-spot network for hyperspectral anomaly detection," *Remote Sensing*, vol. 16, no. 16, p. 3036, 2024.
- [8] R. Zhao and L. Zhang, "Gsead: graphical scoring estimation for hyperspectral anomaly detection," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 10, no. 2, pp. 725–739, 2016.
- [9] Y. Zhang, J. Wang, Y. Chen, H. Yu, and T. Qin, "Adaptive Memory Networks with Self-supervised Learning for Unsupervised Anomaly Detection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 4347, no. c, pp. 1–13, 2022.
- [10] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "Simplenet: A simple network for image anomaly detection and localization," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 20402–20411.
- [11] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le, "A unified model for multi-class anomaly detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4571–4584, 2022.
- [12] Z. Zhong, K. Yang, Z. Yu, Y. Shi, and C. P. Chen, "Towards efficient anomaly detection using memory broad learning system," in *2023 9th International Conference on Control Science and Systems Engineering (ICCSSE)*. IEEE, 2023, pp. 252–257.
- [13] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," in *International Conference on Learning Representations*, 2021.
- [14] A. Garg, W. Zhang, J. Samaran, R. Savitha, and C.-S. Foo, "An evaluation of anomaly detection and diagnosis in multivariate time series," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 6, pp. 2508–2517, 2021.
- [15] S. Kim, K. Choi, H.-S. Choi, B. Lee, and S. Yoon, "Towards a Rigorous Evaluation of Time-Series Anomaly Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, pp. 7194–7201, 2022.
- [16] K. Doshi, S. Abudalou, and Y. Yilmaz, "Reward Once, Penalize Once: Rectifying Time Series Anomaly Detection," in *2022 International Joint Conference on Neural Networks (IJCNN)*, 2022, pp. 1–8.
- [17] R. Wu and E. J. Keogh, "Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the Illusion of Progress (Extended Abstract)," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. Kuala Lumpur, Malaysia: IEEE, 2022, pp. 1479–1480.
- [18] R. Zhao, B. Du, L. Zhang, and L. Zhang, "A robust background regression based score estimation algorithm for hyperspectral anomaly detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 122, pp. 126–144, 2016.
- [19] J. Audibert, P. Michiardi, F. Guyard, S. Marti, and M. A. Zuluaga, "USAD: UnSupervised Anomaly Detection on Multivariate Time Series," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Virtual Event CA USA: ACM, 2020, pp. 3395–3404.
- [20] Y. Yang, C. Zhang, T. Zhou, Q. Wen, and L. Sun, "Dcdetector: Dual attention contrastive representation learning for time series anomaly detection," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023, pp. 3033–3045.
- [21] S. Tuli, G. Casale, and N. R. Jennings, "Tranad: deep transformer networks for anomaly detection in multivari-

- ate time series data,” *Proc. VLDB Endow.*, vol. 15, no. 6, p. 1201–1214, Feb. 2022.
- [22] R. Zhao, B. Du, and L. Zhang, “Hyperspectral anomaly detection via a sparsity score estimation framework,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 6, pp. 3208–3222, 2017.
- [23] R. Zhao, B. Du, L. Zhang, and L. Zhang, “Beyond background feature extraction: An anomaly detection algorithm inspired by slowly varying signal analysis,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 3, pp. 1757–1774, 2016.
- [24] R. Zhao, B. Du, and L. Zhang, “A robust nonlinear hyperspectral anomaly detection approach,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 4, pp. 1227–1234, 2014.
- [25] R. Zhao, Z. Yang, X. Meng, and F. Shao, “A novel fully convolutional auto-encoder based on dual clustering and latent feature adversarial consistency for hyperspectral anomaly detection,” *Remote Sensing*, vol. 16, no. 4, p. 717, 2024.
- [26] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft, “Deep one-class classification,” in *International conference on machine learning*. PMLR, 2018, pp. 4393–4402.
- [27] H. Xu, G. Pang, Y. Wang, and Y. Wang, “Deep isolation forest for anomaly detection,” *IEEE Transactions on Knowledge and Data Engineering*, 2023.
- [28] B. Du, R. Zhao, L. Zhang, and L. Zhang, “A spectral-spatial based local summation anomaly detection method for hyperspectral images,” *Signal Processing*, vol. 124, pp. 115–131, 2016.
- [29] Z. Tian, M. Zhuo, L. Liu, J. Chen, and S. Zhou, “Anomaly detection using spatial and temporal information in multivariate time series,” *Scientific Reports*, vol. 13, no. 1, p. 4400, 2023.
- [30] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, “Timesnet: Temporal 2d-variation modeling for general time series analysis,” in *The eleventh international conference on learning representations*, 2022.
- [31] F. Zeng, M. Chen, C. Qian, Y. Wang, Y. Zhou, and W. Tang, “Multivariate time series anomaly detection with adversarial transformer architecture in the Internet of Things,” *Future Generation Computer Systems*, vol. 144, pp. 244–255, 2023.
- [32] A.-H. Shin, S. T. Kim, and G.-M. Park, “Time Series Anomaly Detection Using Transformer-Based GAN With Two-Step Masking,” *IEEE Access*, vol. 11, pp. 74 035–74 047, 2023.
- [33] B. Du, X. Sun, J. Ye, K. Cheng, J. Wang, and L. Sun, “GAN-Based Anomaly Detection for Multivariate Time Series Using Polluted Training Set,” *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.
- [34] J. Kim, H. Kang, and P. Kang, “Time-series anomaly detection with stacked Transformer representations and 1D convolutional network,” *Engineering Applications of Artificial Intelligence*, vol. 120, p. 105964, 2023.
- [35] H. Xu, Y. Wang, S. Jian, Q. Liao, Y. Wang, and G. Pang, “Calibrated one-class classification for unsupervised time series anomaly detection,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 11, pp. 5723–5736, 2024.
- [36] C. U. Carmona, F.-X. Aubet, V. Flunkert, and J. Gasthaus, “Neural contextual anomaly detection for time series,” in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, L. D. Raedt, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2022, pp. 2843–2851, main Track.
- [37] H. Zhou, K. Yu, X. Zhang, G. Wu, and A. Yazidi, “Contrastive autoencoder for anomaly detection in multivariate time series,” *Information Sciences*, vol. 610, pp. 266–280, 2022.
- [38] T. Pranavan, T. Sim, A. Ambikapathi, and S. Ramasamy, “Contrastive predictive coding for Anomaly Detection in Multi-variate Time Series Data,” *arXiv.org*, 2022.
- [39] C. Wang, Z. Zhuang, Q. Qi, J. Wang, X. Wang, H. Sun, and J. Liao, “Drift doesn’t matter: Dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [40] Y. Chen, C. Zhang, M. Ma, Y. Liu, R. Ding, B. Li, S. He, S. Rajmohan, Q. Lin, and D. Zhang, “Imdiffusion: Imputed diffusion models for multivariate time series anomaly detection,” *Proc. VLDB Endow.*, vol. 17, no. 3, p. 359–372, Nov. 2023.
- [41] T. Zhou, P. Niu, X. Wang, L. Sun, and R. Jin, “One Fits All: Power General Time Series Analysis by Pretrained LM,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 43 322–43 355, 2023.
- [42] S. Mauceri, J. Sweeney, M. Nicolau, and J. McDermott, “Dissimilarity-preserving representation learning for one-class time series classification,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 10, pp. 13 951–13 962, 2024.
- [43] T. Kim, J. Kim, Y. Tae, C. Park, J.-H. Choi, and J. Choo, “Reversible instance normalization for accurate time-series forecasting against distribution shift,” in *International Conference on Learning Representations*, 2021.
- [44] C. Chang, W.-Y. Wang, W.-C. Peng, and T.-F. Chen, “Llm4ts: Aligning pre-trained llms as data-efficient time-series forecasters,” *arXiv preprint arXiv:2308.08469*, 2024.
- [45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
- [46] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan *et al.*, “Time-lm: Time series forecasting by reprogramming large language models,” *arXiv preprint arXiv:2310.01728*, 2023.
- [47] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

- [48] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [49] X. Chen and K. He, “Exploring simple siamese representation learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 15 750–15 758.
- [50] A. Huet, J. M. Navarro, and D. Rossi, “Local Evaluation of Time Series Anomaly Detection Algorithms,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 635–645.
- [51] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh, “The ucr time series archive,” *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 6, pp. 1293–1305, 2019.
- [52] J. Paparrizos, P. Boniol, T. Palpanas, R. S. Tsay, A. Elmore, and M. J. Franklin, “Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection,” *Proceedings of the VLDB Endowment*, vol. 15, no. 11, pp. 2774–2787, 2022.
- [53] C.-Y. A. Lai, F.-K. Sun, Z. Gao, J. H. Lang, and D. Boning, “Nominality Score Conditioned Time Series Anomaly Detection by Point/Sequential Reconstruction,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 76 637–76 655, 2023.
- [54] D. Kim, S. Park, and J. Choo, “When model meets new normals: Test-time adaptation for unsupervised time-series anomaly detection,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 12, pp. 13 113–13 121, Mar. 2024.
- [55] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, “A time series is worth 64 words: Long-term forecasting with transformers,” in *The Eleventh International Conference on Learning Representations*, 2023, pp. 1–22.
- [56] K. Yi, Q. Zhang, W. Fan, S. Wang, P. Wang, H. He, N. An, D. Lian, L. Cao, and Z. Niu, “Frequency-domain mlps are more effective learners in time series forecasting,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [57] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, “Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding,” in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 387–395.
- [58] D. Entekhabi, E. G. Njoku, P. E. O’neill, K. H. Kellogg, W. T. Crow, W. N. Edelstein, J. K. Entin, S. D. Goodman, T. J. Jackson, J. Johnson *et al.*, “The soil moisture active passive (smap) mission,” *Proceedings of the IEEE*, vol. 98, no. 5, pp. 704–716, 2010.
- [59] A. Abdulaal, Z. Liu, and T. Lancewicki, “Practical approach to asynchronous multivariate time series anomaly detection and localization,” in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 2485–2494.
- [60] R. Wu and E. J. Keogh, “Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 3, pp. 2421–2429, 2023.
- [61] S. Sørbo and M. Ruocco, “Navigating the metric maze: a taxonomy of evaluation metrics for anomaly detection in time series,” *Data Min. Knowl. Discov.*, vol. 38, no. 3, p. 1027–1068, Nov. 2023.
- [62] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, “Bootstrap your own latent-a new approach to self-supervised learning,” *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.
- [63] R. Wang, C. Liu, X. Mou, K. Gao, X. Guo, P. Liu, T. Wo, and X. Liu, “Deep contrastive one-class time series anomaly detection,” in *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*. SIAM, 2023, pp. 694–702.



Zhijie Zhong received the B.S. degree in 2022 from the Harbin Engineering University, Harbin, China and he is currently working toward the Ph.D. degree with the South China University of Technology, Guangzhou, China. His research interests include data mining, machine learning, time series analysis, anomaly detection, and large language model (LLM).



Zhiwen Yu (S’06-M’08-SM’14) is a Professor in School of Computer Science and Engineering, South China University of Technology, China. He received the Ph.D. degree from the City University of Hong Kong, Hong Kong, in 2008. Dr. Yu has authored or coauthored more than 200 refereed journal articles and international conference papers, including more than 70 articles in the journals of IEEE Transactions. His google citation is more than 10000, and h-index is 44. He is an Associate Editor of the IEEE Transactions on systems, man, and cybernetics: systems. He is a senior member of IEEE and ACM, a Member of the Council of China Computer Federation (CCF).



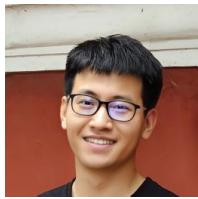
Xing Xi received the B.S. degree in 2020 from the Guangdong Baiyun University and the M.A. degree from the Guangdong University of Technology, Guangzhou, China. Current, he is currently working toward the D.E degree with the South China University of Technology. His research interests include object detection, open world and vocabulary object detection.



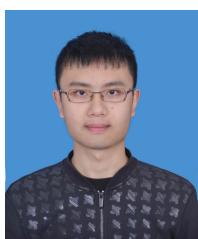
Yue Xu received the B.S. degree in 2025 from the South China University of Technology, Guangzhou, China and he is currently working toward the M.A. degree with the South China University of Technology, Guangzhou, China. His research interests include adversarial attack, continual learning and anomaly detection.



Wenming Cao (S'16) received M.Sc. degree from the School of Automation, Huazhong University of Science and Technology (HUST), Wuhan, China, 2015, and Ph.D. degree at the Department of Computer Science, City University of Hong Kong, 2019. He has been a Postdoctoral Fellow at the University of Hong Kong from 2019 to 2021. He is an associate professor with the Department of Information and Computing Science, Chongqing Jiaotong University. His research interests include data mining and machine learning.



Yiyuan Yang is a DPhil student at the Department of Computer Science, University of Oxford, United Kingdom. He focuses on the field of intelligent sensing systems, time series, spatiotemporal data mining, anomaly detection, and generative models. He previously studied at the Department of Automation at Tsinghua University and interned at the Alibaba DAMO Academy and Huawei Noah's Ark Lab.



Kaixiang Yang (M'21) received the B.S. degree and M.S. degree from the University of Electronic Science and Technology of China and Harbin Institute of Technology, China, in 2012 and 2015, respectively, and the Ph.D. degree from the School of Computer Science and Engineering, South China University of Technology, China, in 2020.

He has been a Research Engineer with the 7th Research Institute, China Electronics Technology Group Corporation, Guangzhou, China, from 2015 to 2017, and has been a Postdoctoral Researcher with Zhejiang University from 2020 to 2021. He is now with the School of Computer Science and Engineering, South China University of Technology. His research interests include pattern recognition, machine learning, and industrial data intelligence.



Jane You received the BEng degree in electronics engineering from Xi'an Jiaotong University, Xi'an, China, in 1986, and the PhD degree in computer science from La Trobe University, Melbourne, VIC, Australia, in 1992. She was a Lecturer with the University of South Australia, Adelaide SA, Australia, and a Senior Lecturer with Griffith University, Nathan, QLD, Australia, from 1993 to 2002. She is currently a Full Professor with The Hong Kong Polytechnic University, Hong Kong. Her current research interests include image processing, pattern recognition, medical imaging, biometrics computing, multimedia systems, and data mining.