Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy

Abstract

Unsupervised detection of anomaly points in time series is a challenging problem that requires deriving a distinguishable criterion. Previous approaches have mainly focused on learning pointwise representations or pairwise associations, yet neither is sufficient to reason about the intricate dynamics. Recently, Transformers have demonstrated great power in unifying pointwise representations and pairwise associations. This study finds that the self-attention weight distribution of each time point can embody rich associations with the entire series. The rarity of anomalies makes it extremely difficult to build nontrivial associations from abnormal points to the whole series. Therefore, the anomalies' associations should mainly concentrate on their adjacent time points, which implies an association-based criterion inherently distinguishable between normal and abnormal points. This paper highlights this through the *Association Discrepancy*. Technically, the proposed *Anomaly Transformer* utilizes an *Anomaly-Attention* mechanism to calculate the association discrepancy. A minimax strategy is employed to amplify the normal-abnormal distinguishability of the association discrepancy. The Anomaly Transformer achieves state-of-the-art results on six unsupervised time series anomaly detection benchmarks across three applications: service monitoring, space & earth exploration, and water treatment.

Keywords: Anomaly detection, Anomaly Transformer, Association discrepancy.

1 Introduction

Unsupervised detection of time series outliers is a challenging problem. Fault detection from large-scale system monitoring data can be simplified to finding abnormal time points from time series. But anomalies are usually rare and hidden by a large number of normal points, which makes data annotation difficult and expensive. Various classical anomaly detection methods do not consider temporal information and are difficult to generalize to unseen real scenes.

Recent deep models have demonstrated their powerful learning capabilities. One major approach focuses on recurrent neural networks learning pointwise representations and performs self-supervision through reconstruction or autoregressive tasks [13]. The basis for anomaly detection is the pointwise reconstruction or prediction error. However, for complex temporal patterns, pointwise representations contain limited information, and they are dominated by normal time points, making it difficult to distinguish anomalies. Furthermore, reconstruction or prediction is performed on a per-point basis and cannot provide a comprehensive description of the temporal context.

Another major approach is based on explicit correlation modeling for anomaly detection. For example, graph neural networks are used to construct such correlations [3]. Although graph neural networks have stronger expressive power, the learned graphs are still limited to individual time points. Methods based on subsequence calculation compute similarities between subsequences to detect anomalies. However, these methods fail to capture fine-grained temporal correlations between each time point and the entire sequence when exploring a broader temporal context.

This paper proposes an Anomaly Transformer based on the Anomaly-Attention mechanism, which innovatively leverages differences in correlated observations. Transformers are applied to time series, and the time correlation of each time point can be obtained from the self-attention graph. The self-attention graph represents the distribution of correlation weights along the temporal dimension for all time points. The correlation distribution of each time point can provide a more informative description of the temporal context, indicating dynamic patterns such as cycles or trends in the time series. The correlation distribution, referred to as series-association in the paper, can be discovered by Transformers from the original sequence.

Anomaly Transformer achieves state-of-the-art results on six unsupervised time series anomaly detection benchmarks across three application scenarios: service detection, space & Earth exploration, and water treatment, with extensive ablations and insightful case studies.

2 Related works

Unsupervised time series anomaly detection is extremely challenging in practice. Models should learn representations of information from complex temporal dynamics through unsupervised tasks. Nonetheless, it should also lead to a distinguishing criterion that can detect rare anomalies from a large number of normal time points.

2.1 Classical anomaly detection methods

Various classical anomaly detection methods provide many unsupervised paradigms, such as the density estimation method proposed in the local anomaly factor (LOF) [2], Cluster-based methods proposed in one-class SVM (OC-SVM) [11] and SVDD [14]. These classical methods do not consider temporal information and are difficult to generalize to unseen real scenes.

2.2 Learning pointwise representations

Thanks to the representation learning capabilities of neural networks, recent deep models [13] have achieved remarkable performance. One major approach focuses on learning pointwise representations through well-designed recurrent networks and performs self-supervision through reconstruction or autoregressive tasks. Here, a natural and practical criterion for anomaly detection is the pointwise reconstruction or prediction error. However, due to the rarity of anomalies, pointwise representations contain limited information for complex temporal patterns, and they are dominated by normal time points, making it difficult to distinguish anomalies. Furthermore, reconstruction or prediction errors are computed on a per-point basis and cannot provide a comprehensive description of the temporal context.

2.3 Explicit association modeling

Another major approach is based on explicit correlation modeling for anomaly detection. Vector autoregressive models and state space models belong to this category. By representing time series at different time points as vertices and detecting anomalies through random walks, graphs are also used to explicitly capture correlations [3]. Typically, these classical methods have difficulty learning information representation and modeling fine-grained correlations. In recent years, graph neural networks (GNNs) have been used to learn dynamic graphs between multiple variables in multivariate time series [17]. Although they have stronger expressive power, the learned graphs are still limited to individual time points, which is not sufficient for complex temporal patterns.

3 Method

3.1 Overview

Given the limitation of Transformers [15] for anomaly detection, this paper renovate the vanilla architecture to the Anomaly Transformer (Figure 1) with an Anomaly-Attention mechanism.Anomaly-Attention (left) models the prior-association and series-association simultaneously. In addition to the reconstruction loss, our model is also optimized by the minimax strategy with a specially-designed stop-gradient mechanism (gray arrows) to constrain the prior- and series-associations for more distinguishable association discrepancy.At the minimize phase, the prior-association minimizes the Association Discrepancy within the distribution family derived by Gaussian kernel. At the maximize phase, the series-association maximizes the Association Discrepancy under the reconstruction loss.

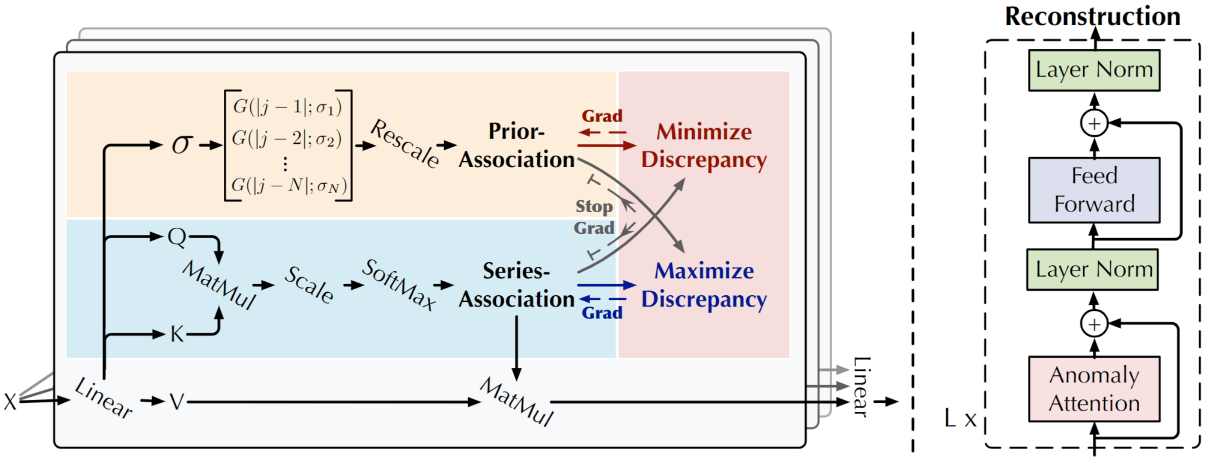


Figure 1. Anomaly Transformer

3.2 Feature extraction

Anomaly Transformer is characterized by stacking the Anomaly-Attention blocks and feed-forward layers alternately. This stacking structure is conducive to learning underlying associations from deep multi-level features. Suppose the model contains L layers with length-N input time series $X \in \mathbb{R}^{N \times d}$.The overall equations of the l-th layer are formalized as:

$$\begin{aligned} Z^l &= \text{Layer} - \text{Norm}\left(\text{Anomaly} - \text{Attention}\left(X^{l-1}\right) + X^{l-1}\right) \\ X^l &= \text{Layer} - \text{Norm}\left(\text{Feed} - \text{Forward}\left(Z^l\right) + Z^l\right) \end{aligned} \tag{1}$$

where $X^l \in \mathbb{R}^{N \times d_{\text{model}}}$, $l \in \{1, ..., L\}$ denotes the output of the l-th layer with d_{model} channels. The initial input $X^0 = \text{Embedding}\,(X)$ represents the embedded raw series. $Z^l \in \mathbb{R}^{N \times d_{\text{model}}}$ is the l-th layer's hidden representation. Anomaly-Attention $(\cdot)$ is to compute the association discrepancy.

Note that the single-branch self-attention mechanism [15] cannot model the prior-association and series-association simultaneously. This paper proposes the Anomaly-Attention with a two-branch structure (Figure 1). In the prior-association, a learnable Gaussian kernel is used to calculate the prior based on the relative temporal distance. This design takes advantage of the unimodal property of the Gaussian kernel, allowing it to focus more on the neighboring temporal context. Additionally, a learnable scale parameter is used to adapt the prior-associations to different time series patterns, including varying lengths of anomaly segments.The series-association branch learns associations from raw series data, enabling it to dynamically identify the most effective associations. It is important to note that both approaches maintain temporal dependencies at each time point, providing more informative representations compared to point-wise representations. They capture the adjacent-concentration prior and learned associations, respectively, which should help distinguish between normal and abnormal patterns.The Anomaly-Attention in the l-th layer is:

$$\begin{aligned} \text{Initialization :}&Q, K, V, \sigma = X^{l-1}W_Q^l, X^{l-1}W_K^l, X^{l-1}W_V^l, X^{l-1}W_\sigma^l \\ \text{Prior - Association :}&P^l = \text{Rescale}\left(\left[\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)\right]_{i,j\in\{1,\ldots,N\}}\right) \\ \text{Series - Association :}&S^l = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_{\text{model}}}}\right) \\ \text{Reconstruction: }&Z^l = S^l V, \end{aligned} \tag{2}$$

where $Q, K, V \in \mathbb{R}^{N \times d_{\text{model}}}$, $\sigma \in \mathbb{R}^{N \times 1}$ represent the query, key, value of self-attention and the learned scale respectively. $W_Q^l, W_K^l, W_V^l \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$, $W_\sigma^l \in \mathbb{R}^{d_{\text{model}} \times 1}$ represent the parameter matrices for Q, K, V, σ in the l-th layer respectively. Prior-association $P^l \in \mathbb{R}^{N \times N}$ is generated based on the learned scale $\sigma \in \mathbb{R}^{N \times 1}$ and the $i-$th element σ_i corresponds to the i-th time point. Concretely, for the i-th time point, its association weight to the j-th point is calculated by the Gaussian kernel $G\left(|j-i|\,;\sigma_i\right) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)$ w.r.t. the distance $|j-i|$. Further, this paper uses Rescale $(\cdot)$ to transform the association weights to discrete distributions P^l by dividing the row sum $S^l \in \mathbb{R}^{N \times N}$ denotes the series-associations. Softmax $(\cdot)$ normalizes the attention map along the last dimension, and each row of S^l forms a discrete distribution. $\hat{Z}^l \in \mathbb{R}^{N \times d_{\text{model}}}$ is the hidden representation after the Anomaly-Attention in the l-th layer. this paper uses Anomaly-Attention $(\cdot)$ to summarize Equation 2. In the multi-head version, the learned scale is $\sigma \in \mathbb{R}^{N\times h}$ for h heads. $Q_m, K_m, V_m \in \mathbb{R}^{N\times \frac{d_{\text{model}}}{h}}$ denote the query, key and value of the m-th head respectively. The block concatenates the outputs $\{\hat{Z}^l \in \mathbb{R}^{N\times \frac{d_{\text{model}}}{h}}\}_{1\leq m\leq h}$ from multiple heads and gets the final result $\hat{Z}^l \in \mathbb{R}^{N\times d_{\text{model}}}$.

3.3 Association discrepancy

This paper formalizes the *Association Discrepancy* as the symmetrized KL divergence between prior- and series-associations, which represents the information gain between these two distributions [9]. The authors average the association discrepancy from multiple layers to combine the associations from multi-level features into a more informative measure as:

$$\text{AssDis}(P, S; X) = \left[\frac{1}{L}\sum_{l=1}^{L}\left(\text{KL}\left(P^l_{i,:}||S^l_{i,:}\right) + \text{KL}\left(S^l_{i,:}||P^l_{i,:}\right)\right)\right]_{i=1,\ldots,N} \tag{3}$$

where $\text{KL}\,(\cdot||\cdot)$ is the KL divergence computed between two discrete distributions corresponding to every row of P^l and S^l. $\text{AssDis}(P, S; X) \in \mathbb{R}^{N\times 1}$ is the point-wise association discrepancy of X with respect to prior-association P and series-association S from multiple layers. The i-th element of AssDis corresponds to the i-th time point of X . From previous observation, anomalies will present smaller $\text{AssDis}(P, S; X)$ than normal time points, which makes AssDis inherently distinguishable.

3.4 Loss

This paper utilizes the reconstruction loss as a means of optimizing our model, since it is an unsupervised task. By doing so, the reconstruction loss guides the series-association in identifying the most informative associations. This paper also incorporates an additional loss to further accentuate the difference between normal and abnormal time points, thereby increasing the association discrepancy. As the prior-association exhibits unimodal properties, the discrepancy loss directs the series-association to prioritize non-adjacent areas, which in turn makes anomaly detection more challenging while simultaneously rendering anomalies more identifiable.The loss function for input series $X \in \mathbb{R}^{N\times d}$ is formalized as:

$$L_{\text{Total}}\left(\hat{X}, P, S, \lambda; X\right) = ||X - \hat{X}||_{\text{F}}^2 - ||\lambda \times \text{AssDis}(P, S; X)||_1 \tag{4}$$

where $\hat{X} \in \mathbb{R}^{N\times d}$ denotes the reconstruction of X. $||\cdot||_{\text{F}}, ||\cdot||_k$ indicate the Frobenius and k-norm. λ is to trade off the loss terms. When $\lambda > 0$, the optimization is to enlarge the association discrepancy. A minimax strategy is proposed to make the association discrepancy more distinguishable.

Note that directly maximizing the association discrepancy will extremely reduce the scale parameter of the Gaussian kernel [9], making the prior-association meaningless. Towards a better control of association learning, this paper proposes a minimax strategy (Figure 2).

Concretely, for the minimize phase, this paper drives the prior-association P^l to approximate the series-association S^l that is learned from raw series. This process will make the prior-association adapt to various temporal patterns. For the maximize phase, this paper optimizes the series-association to enlarge the association discrepancy. This process forces the series-association to pay more attention to the non-adjacent horizon.

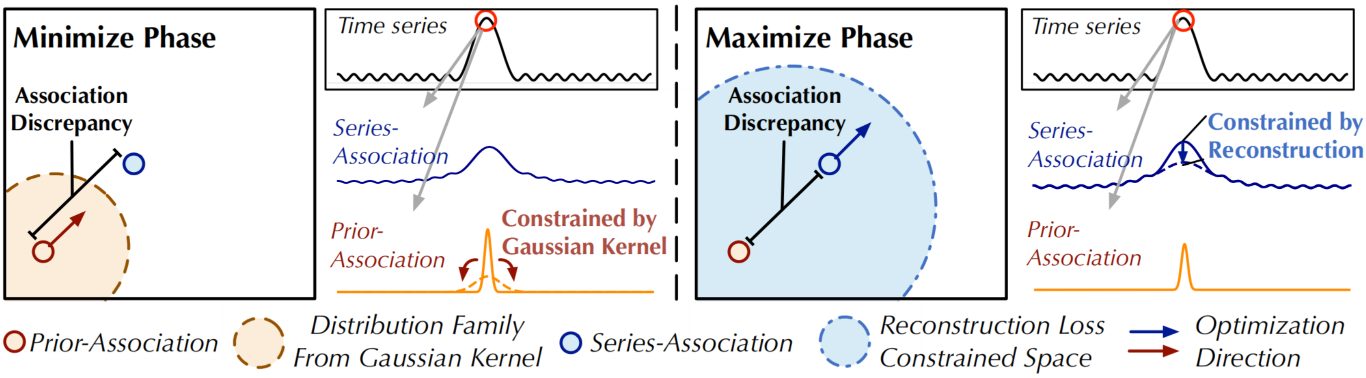


Figure 2. Minimax association learning. At the minimize phase, the prior-association minimizes the Association Discrepancy within the distribution family derived by Gaussian kernel. At the maximize phase, the series-association maximizes the Association Discrepancy under the reconstruction loss.

Thus, integrating the reconstruction loss, the loss functions of two phases are:

$$\begin{aligned}\text{Minimize Phase: } & L_{\text{Total}}(\hat{X}, P, S_{\text{detach}}, -\lambda; X) \\ \text{Maximize Phase: } & L_{\text{Total}}(\hat{X}, P_{\text{detach}}, S, \lambda; X)\end{aligned} \tag{5}$$

where $\lambda > 0$ and $*_{\text{detach}}$ means to stop the gradient backpropagation of the association (Figure 1). As P approximates S_{detach} in the minimize phase, the maximize phase will conduct a stronger constraint to the series-association, forcing the time points to pay more attention to the non-adjacent area. Under the reconstruction loss, this is much harder for anomalies to achieve than normal time points, thereby amplifying the normal-abnormal distinguishability of the association discrepancy.

3.5 Association-based Anomaly Criterion

this paper incorporates the normalized association discrepancy to the reconstruction criterion, which will take the benefits of both temporal representation and the distinguishable association discrepancy. The final anomaly score of $X \in \mathbb{R}^{N \times d}$ is shown as follows:

$$\text{AnomalyScore}\,(X) = \text{Softmax}\,(-\text{AssDis}(P, S; X)) \odot \left[||X_{i,:} - \hat{X}_{i,:}||_2^2 \right]_{i=1,\ldots,N} \tag{6}$$

where $\odot$ is the element-wise multiplication. $\text{AnomalyScore}(X) \in \mathbb{R}^{N \times 1}$ denotes the point-wise anomaly criterion of X . Towards a better reconstruction, anomalies usually decrease the association discrepancy, which will still derive a higher anomaly score. Thus, this design can make the reconstruction error and the association discrepancy collaborate to improve detection performance.

4 Implementation details

The implementation experiment was based on five experimental datasets provided by the authors:

(1) SMD(Server Machine Dataset [13]), a 5-week long dataset with 38 dimensions collected by a large Internet company;

(2) PSM(Pooled Server Metrics [1]), a 26-dimensional data set collected from multiple application server nodes of eBay;

(3) Both MSL (Mars Science Laboratory rover) and SMAP (Soil Moisture Active Passive satellite) are public datasets from NASA [4] with 55 and 25 dimensions respectively, which contain the telemetry anomaly data derived from the Incident Surprise Anomaly (ISA) reports of spacecraft monitoring systems.

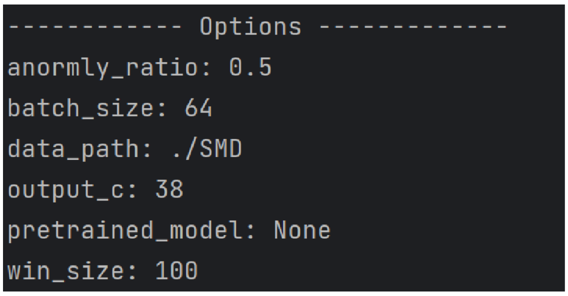
(4) SWaT(Secure Water Treatment [8]), obtained from 51 sensors in a critical infrastructure system under continuous operation.

4.1 Comparing with the released source codes

The author employed the Mean Squared Error (MSE) loss function, a commonly used loss function in computer vision tasks, for measuring the reconstruction error. In contrast, I opted to replace it with the SmoothL1Loss loss function, which has been shown to exhibit better robustness to outliers [7]. This choice allows for a more precise evaluation of the reconstruction quality and contributes to enhancing the overall robustness and performance of the model.

4.2 Experimental environment setup

Following the well-established protocol [12], a non-overlapping sliding window method was used to obtain a set of subsequencing, and the window size was fixed to 100 for all datasets. A time point is labeled as anomalous if its anomaly score(Equation 6), exceeds a certain threshold δ. This threshold δ is determined to ensure that a fraction r of time points in the validation dataset are labeled as anomalous. $r = 0.1\%$ for SWaT, 0.5% for SMD, and 1% for the other datasets were set to obtain the main results. Considering that abnormal time points may indicate the presence of multiple anomalies within a single abnormal segment, the authors employ a commonly used adjustment strategy [16], where all anomalies within a detected anomaly segment are counted as correctly detected. The abnormal transformer consists of three layers, the number of hidden state channels d_{model} is 512, and the number of magnetic heads h is set to 8. For all datasets, I set the hyperparameter λ(Equation 4) to 3 to balance the two parts of the loss function. The authors used ADAM optimizer [5]with an initial learning rate of 10^{-4}. The training process is stopped early within 10 epochs with a batch size of 64. All experiments were performed in Python 3.7 and PyTorch 1.4 [10]with a single NVIDIA GeForce RTX 3050 4GB Laptop GPU.

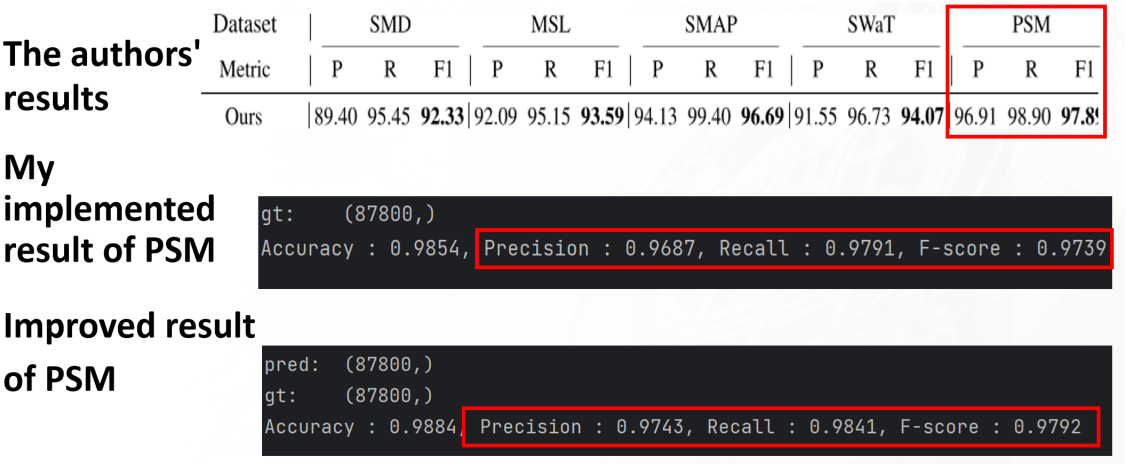


```
------------ Options -------------
anormly_ratio: 0.5
batch_size: 64
data_path: ./SMD
output_c: 38
pretrained_model: None
win_size: 100
```

Figure 3. Partial parameter setting, taking SMD dataset as example.

4.3 Main contributions

The advantage of using the SmoothL1Loss function over the Mean Squared Error (MSE) loss function originally used in the paper is that it penalizes outliers less, thus making the model more robust. Although the ultimate goal is to detect errors, most detection errors in reconstruction error occur when normal points are mistakenly classified as anomalous, given the guarantee of association differences. Therefore, my consideration is actually to reduce the sensitivity of reconstruction error to anomalous points. As my implemented results are not identical to the authors', I mainly compare my implemented and improved results. While the performance on the four datasets does not differ significantly from that obtained with MSE loss, there is a slight improvement observed(Figure 4).



Dataset	SMD			MSL			SMAP			SWaT			PSM		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Ours	89.40	95.45	**92.33**	92.09	95.15	**93.59**	94.13	99.40	**96.69**	91.55	96.73	**94.07**	96.91	98.90	**97.8**

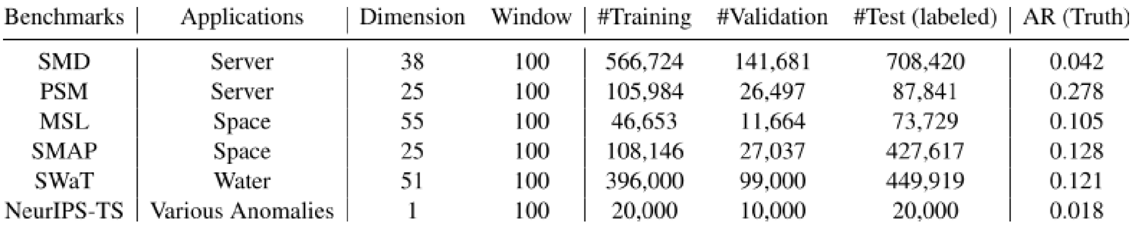
```
gt:     (87800,)
Accuracy : 0.9854, Precision : 0.9687, Recall : 0.9791, F-score : 0.9739
```

```
pred:  (87800,)
gt:    (87800,)
Accuracy : 0.9884, Precision : 0.9743, Recall : 0.9841, F-score : 0.9792
```

Figure 4. Comparison of experimental results, taking PSM dataset as example.

5 Results and analysis

this paper uses six experimental datasets(Figure 5), which contain three real world applications, including server, low altitude detection, water quality detection. The last one is the dataset proposed by a paper published in Neurls2021 [6], which mainly detects whether the algorithm has multiple time series pattern recognition ability.



Benchmarks	Applications	Dimension	Window	#Training	#Validation	#Test (labeled)	AR (Truth)
SMD	Server	38	100	566,724	141,681	708,420	0.042
PSM	Server	25	100	105,984	26,497	87,841	0.278
MSL	Space	55	100	46,653	11,664	73,729	0.105
SMAP	Space	25	100	108,146	27,037	427,617	0.128
SWaT	Water	51	100	396,000	99,000	449,919	0.121
NeurIPS-TS	Various Anomalies	1	100	20,000	10,000	20,000	0.018

Figure 5. Six benchmarks for three practical applications.

Firstly, The experimental results on real world datasets show that Anomaly Transformer achieves SOTA on five datasets, and has a great improvement over the previous 18 baselines in terms of precision, recall and F1score(Figure 6). Anomaly Transformer achieves the consistent state-of-theart on all benchmarks.

Dataset	SMD			MSL			SMAP			SWaT			PSM		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
OCSVM	44.34	76.72	56.19	59.78	86.87	70.82	53.85	59.07	56.34	45.39	49.22	47.23	62.75	80.89	70.67
IsolationForest	42.31	73.29	53.64	53.94	86.54	66.45	52.39	59.07	55.53	49.29	44.95	47.02	76.09	92.45	83.48
LOF	56.34	39.86	46.68	47.72	85.25	61.18	58.93	56.33	57.60	72.15	65.43	68.62	57.89	90.49	70.61
Deep-SVDD	78.54	79.67	79.10	91.92	76.63	83.58	89.93	56.02	69.04	80.42	84.45	82.39	95.41	86.49	90.73
DAGMM	67.30	49.89	57.30	89.60	63.93	74.62	86.45	56.73	68.51	89.92	57.84	70.40	93.49	70.03	80.08
MMPCACD	71.20	79.28	75.02	81.42	61.31	69.95	88.61	75.84	81.73	82.52	68.29	74.73	76.26	78.35	77.29
VAR	78.35	70.26	74.08	74.68	81.42	77.90	81.38	53.88	64.83	81.59	60.29	69.34	90.71	83.82	87.13
LSTM	78.55	85.28	81.78	85.45	82.50	83.95	89.41	78.13	83.39	86.15	83.27	84.69	76.93	89.64	82.80
CL-MPPCA	82.36	76.07	79.09	73.71	88.54	80.44	86.13	63.16	72.88	76.78	81.50	79.07	56.02	99.93	71.80
ITAD	86.22	73.71	79.48	69.44	84.09	76.07	82.42	66.89	73.85	63.13	52.08	57.08	72.80	64.02	68.13
LSTM-VAE	75.76	90.08	82.30	85.49	79.94	82.62	92.20	67.75	78.10	76.00	89.50	82.20	73.62	89.92	80.96
BeatGAN	72.90	84.09	78.10	89.75	85.42	87.53	92.38	55.85	69.61	64.01	87.46	73.92	90.30	93.84	92.04
OmniAnomaly	83.68	86.82	85.22	89.02	86.37	87.67	92.49	81.99	86.92	81.42	84.30	82.83	88.39	74.46	80.83
InterFusion	87.02	85.43	86.22	81.28	92.70	86.62	89.77	88.52	89.14	80.59	85.58	83.01	83.61	83.45	83.52
THOC	79.76	90.95	84.99	88.45	90.97	89.69	92.06	89.34	90.68	83.94	86.36	85.13	88.14	90.99	89.54
Ours	89.40	95.45	**92.33**	92.09	95.15	**93.59**	94.13	99.40	**96.69**	91.55	96.73	**94.07**	96.91	98.90	**97.89**

Figure 6. Quantitative results for Anomaly Transformer in the five datasets. The P, R and F1 represent the precision, recall and F1-score (as %) respectively. F1-score is the harmonic mean of precision and recall. For these three metrics, a higher value indicates a better performance.

Next is the experimental result of NeurIPS-TS benchmark. It mainly divides anomalies into point anomalies and pattern anomalies, which contains these five types of anomalies. Anomaly Transformer can still achieve SOTA performance(Figure 7).It can be seen that from the graph that it has the highest F1score. This verifies the effectiveness of anomaly Transformer on various anomalies.

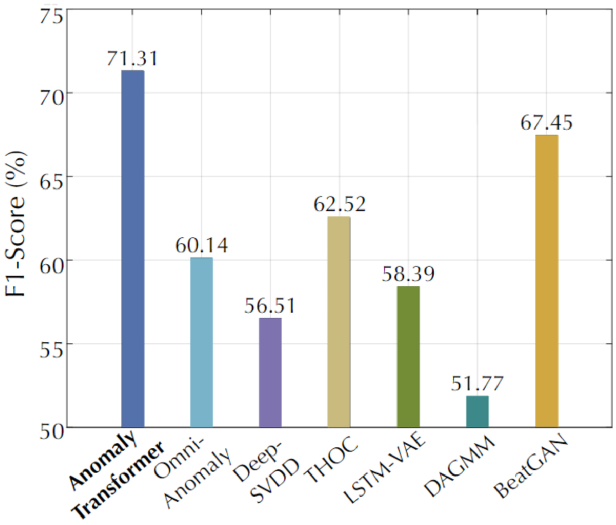
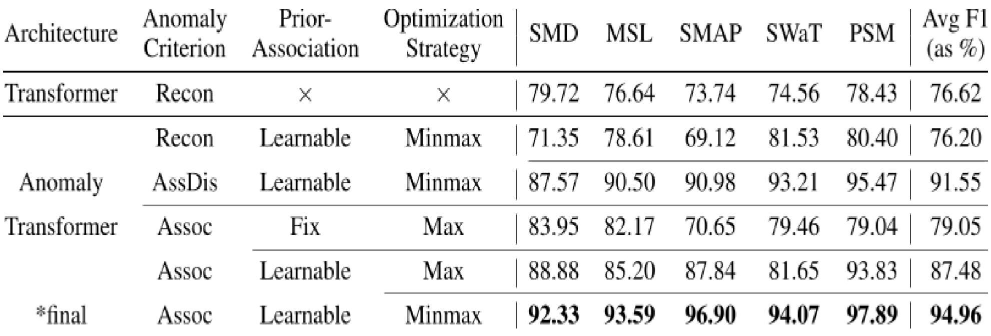


Figure 7. Results for NeurIPS-TS.

The next step was to conduct ablation experiments to further investigate the impact of each part of the model.From the experimental results(Figure 8), it can be seen that each part has improved the model, such as 18 (76.2-94.96) after adding the anomaly criterion, 8 (79-87) after adding PA, and 7 (87-94) after adding the minimax strategy. These verify that each module of the design is effective and necessary.



Architecture	Anomaly Criterion	Prior-Association	Optimization Strategy	SMD	MSL	SMAP	SWaT	PSM	Avg F1 (as %)
Transformer	Recon	×	×	79.72	76.64	73.74	74.56	78.43	76.62
Anomaly Transformer	Recon	Learnable	Minmax	71.35	78.61	69.12	81.53	80.40	76.20
	AssDis	Learnable	Minmax	87.57	90.50	90.98	93.21	95.47	91.55
	Assoc	Fix	Max	83.95	82.17	70.65	79.46	79.04	79.05
	Assoc	Learnable	Max	88.88	85.20	87.84	81.65	93.83	87.48
*final	Assoc	Learnable	Minmax	**92.33**	**93.59**	**96.90**	**94.07**	**97.89**	**94.96**

Figure 8. Ablation results (F1-score) in anomaly criterion, prior-association and optimization strategy. *Recon, AssDis* and *Assoc* mean the pure reconstruction performance, pure association discrepancy and our proposed association-based criterion respectively. *Fix* is to fix *Learnable* scale parameter σ of prior-association as 1.0. *Max* and *Minimax* refer to the strategies for association discrepancy in the maximization (Equation 4) and minimax (Equation 5) way respectively.

This paper also presents visualization experiments for different classes of anomalies(Figure 9). It can be seen that contextual anomaly has a large fluctuation for the reconstruction criterion, and some abnormal points will be hidden in the normal points, while the abnormal criterion based on association has a large gap between the normal and abnormal points, which also recalls the reason for setting up such a criterion before. Compared with the reconstruction criterion, the pattern anomaly in this section can also be well detected. This verifies that the criterion can highlight the anomalies and provide distinct values for normal and abnormal points.

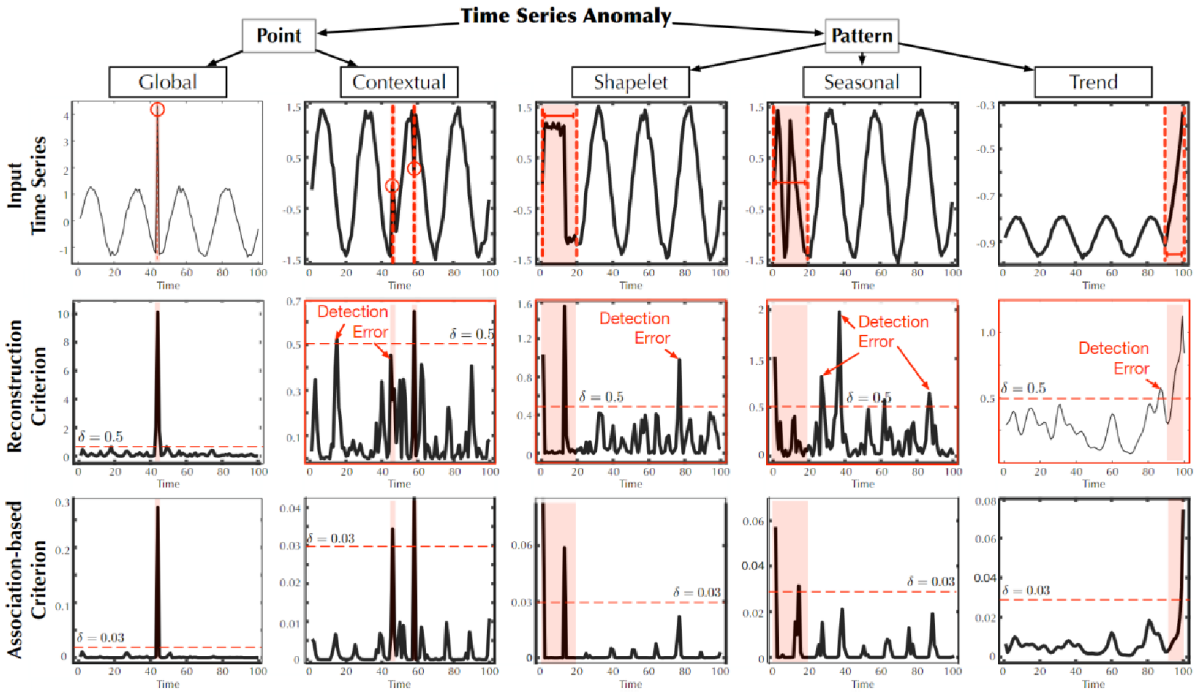
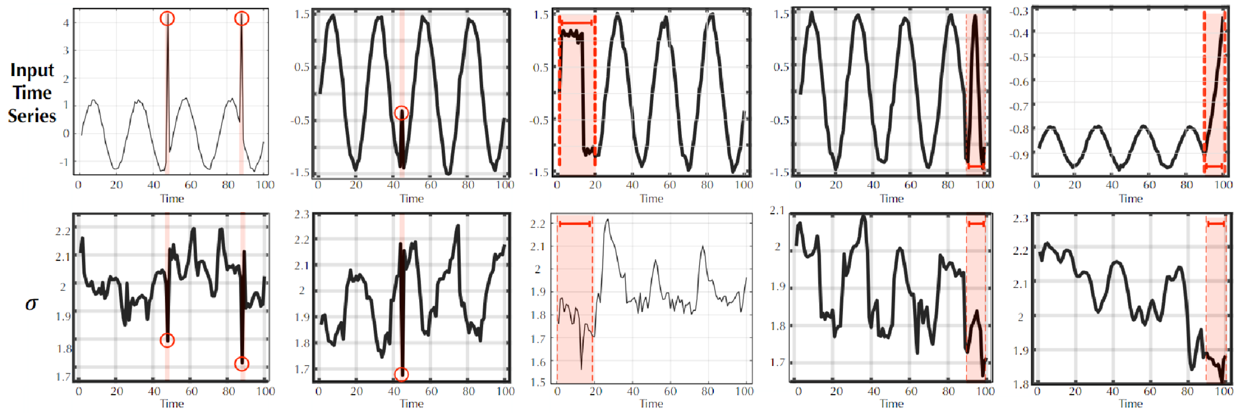


Figure 9. Visualization of different anomaly categories. The authors plot the raw series(first row) from NeurlPS-TS dataset, as well as their corresponding reconstruction (second row) and association-based criteria (third row). The point-wise anomalies are marked by red circles and the pattern-wise anomalies are in red segments. The wrongly detected cases are bounded by red boxes.

In addition, in order to confirm the hypothesis of proximity concentration, the author also did a visualization experiment(Figure 10). It is said that σ learned by PA can reflect the adjacent concentration degree of time series. From the figure, it can be found that σ changes to adapt to various data patterns of time series. In particular, the prior-association of anomalies generally has a smaller σ than normal time points, which matches our adjacent-concentration inductive bias of anomalies.



(a) Point-Global (b) Point-Contextual (c) Pattern-Shapelet (d) Pattern-Seasonal (e) Pattern-Trend

Figure 10. Learned scale parameter σ for different types of anomalies (highlight in red).

Finally, a statistical experiment is conducted to compare adjacent association weights for Abnormal and Normal time points(Figure 11). It can be found that the discrimination between normal and abnormal points is not high (the contrast value basically closed to 1) by only using the reconstruction loss, and the direct maximization of the initial loss function cannot amplify the difference between normal and abnormal points as expected (for example, the contrast value of SMAP still closed to 1), but after adding the minimax strategy, the contrast value is greatly improved.

Dataset	SMD			MSL			SMAP			SWaT			PSM		
Optimization	Recon	Max	Ours	Recon	Max	Ours	Recon	Max	Ours	Recon	Max	Ours	Recon	Max	Ours
Abnormal (%)	1.08	0.95	0.86	1.01	0.65	0.35	1.29	1.18	0.70	1.27	0.89	0.37	1.02	0.56	0.29
Normal (%)	0.94	0.75	0.36	1.00	0.59	0.22	1.23	1.09	0.49	1.18	0.78	0.21	0.99	0.54	0.11
Contrast $\left(\frac{\text{Abnormal}}{\text{Normal}}\right)$	1.15	1.27	**2.39**	1.01	1.10	**1.59**	1.05	1.08	**1.43**	1.08	1.14	**1.76**	1.03	1.04	**2.64**

Figure 11. Results of adjacent association weights for *Abnormal* and *Normal* time points respectively. *Recon, Max* and *Minimax* represent the association learning process that is supervised by reconstruction loss, direct maximization and minimax strategy respectively. A higher contrast value $\left(\frac{\text{Abnormal}}{\text{Normal}}\right)$ indicates a stronger distinguishability between normal and abnormal time points.

6 Conclusion and future work

To sum up, in this paper, the authors introduce Transformers into the unsupervised time series anomaly detection problem. Based on the key observation of association discrepancy, the authors propose Anomaly Transformer, including an Anomaly Attention with the two-branch structure to embody the association discrepancy. A minimax strategy is adopted to further amplify the difference between normal and abnormal time points. Through an exhaustive set of empirical studies, it is proved that Anomaly Transformer achieves the state-of-the-art results, which also proves its rationality. Future work includes theoretical study of Anomoly Transformer in light of classic analysis for autoregression and state space models.

Here are some Disadvantages of this paper. Firstly, this paper does not specify whether pattern-wise anomalies detection also performs better. Secondly, the experiments on the NeurIPS-TS dataset only compare the deep models and do not compare the non-deep model, so whether the performance of this model is better than the classical method in different abnormal patterns remains to be verified.

References

[1] Ahmed Abdulaal, Zhuanghua Liu, and Tomer 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*, pages 2485–2494, 2021.

[2] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and J¨org Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

[3] Haibin Cheng, Pang-Ning Tan, Christopher Potter, and Steven A. Klooster. A robust graph-based algorithm for detection and characterization of anomalies in noisy multivariate time series. In *IEEE international conference on data mining workshops*, pages 349–358, 2008.

[4] Eamonn J. Keogh, Taposh Roy, Naik U, and Agrawal A. Multi-dataset time-series anomaly detection competition. In *Competition of International Conference on Knowledge Discovery & Data Mining*, 2021.

[5] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. Int. Conf. on Learning Representations*, page 6980, 2014.

[6] Kwei-Herng Lai, D. Zha, Junjie Xu, and Yue Zhao. Revisiting time series outlier detection: Definitions and benchmarks. In *Thirty-fifth conference on neural information processing systems datasets and benchmarks track*, 2021.

[7] Chao Liu, Shuai Yu, Min Yu, Baole Wei, Boquan Li, Gang Li, and Weiqing Huang. Adaptive smooth l1 loss: A better way to regress scene texts with extreme aspect ratios. In *IEEE Symposium on Computers and Communications*, pages 1–7, 2021.

[8] Aditya P. Mathur and Nils Ole Tippenhauer. Swat: A water treatment testbed for research and training on ics security. In *international workshop on cyber-physical systems for smart water networks*, pages 31–36, 2016.

[9] Radford M. Neal. Pattern recognition and machine learning. *Advances in neural information processing systems*, pages 366–366, 2007.

[10] Adam Paszke, S. Gross, Francisco Massa, A. Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Z. Lin, N. Gimelshein, L. Antiga, Alban Desmaison, Andreas K¨opf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Proc. Conf. on Neural Information Processing Systems*, page 32, 2019.

[11] B. Sch¨olkopf, John C. Platt, J. Shawe-Taylor, Alex Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.

[12] Lifeng Shen, Zhuocong Li, and James T. Kwok. Timeseries anomaly detection using temporal hierarchical one-class network. *Advances in Neural Information Processing Systems*, 33:13016–13026, 2020.

[13] Ya Su, Y. Zhao, Chenhao Niu, Rong Liu, W. Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2828–2837, 2019.

[14] D. Tax and R. Duin. Support vector data description. *Machine learning*, 54:45–66, 2004.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[16] Haowen Xu, Wenxiao Chen, N. Zhao, Zeyan Li, Jiahao Bu, Zhihan Li, Y. Liu, Y. Zhao, Dan Pei, Yang Feng, Jian Jhen Chen, Zhaogang Wang, and Honglin Qiao. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web application. In *Proceedings of the 2018 world wide web conference*, pages 187–196, 2018.

[17] Hang Zhao, Yujing Wang, Juanyong Duan, Congrui Huang, Defu Cao, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Multivariate time-series anomaly detection via graph attention network. In *IEEE International Conference on Data Mining*, pages 841–850, 2020.