Time Series Anomaly Detection: Transformer-Based Methods Exploration and Comparison

Yongjia Huang

The Hong Kong University of Science and Technology (Guangzhou)

Guangzhou Shi, Guangdong Sheng, China

yhuang181@connect.hkust-gz.edu.cn

ABSTRACT

Time series anomaly detection is a critical task in various industrial applications such as predictive maintenance, fault detection, and cybersecurity. Traditional methods often struggle with the high-dimensional and complex nature of time series data. Transformer-based methods have recently emerged as powerful tools for capturing long-range dependencies and temporal patterns in time series data. This paper explores and compares three transformer-based methods for time series anomaly detection: AnomalyBERT [3], AnomalyTransformer [14], and TranAD [12]. I will delve into their model structures, loss functions, advantages, limitations, applicable scenarios, and inference speeds.

1 INTRODUCTION

The increasing prevalence of time-sensitive data across industries necessitates more effective anomaly detection systems. These systems must not only be accurate but also capable of handling large, complex datasets without extensive manual labeling. The challenge lies in adapting the latest AI technologies, such as transformers, to detect subtle and non-obvious anomalies effectively.

In this paper, I compare three state-of-the-art models in time series anomaly detection: AnomalyTransformer, AnomalyBERT, and TranAD. I explore the advantages and disadvantages of each model, their applicable scenarios, and their effectiveness in detecting different types of anomalies. AnomalyTransformer is notable for its association discrepancy mechanism, which enhances its ability to capture long-term dependencies. AnomalyBERT leverages a self-supervised approach with synthetic outliers to train the model, focusing on improving detection accuracy. TranAD employs a combination of self-conditioning and adversarial training to enhance robustness and efficiency.

Particularly, I delve into comparative model experiments that show varied performance across different datasets. This report aims to investigate and explore the reasons behind these observations, providing insights into the strengths and limitations of transformer-based models in time series anomaly detection.

2 RELATED WORK

2.1 Unsupervised Time Series Anomaly Detection

Unsupervised anomaly detection problems have been handled using various statistical and machine learning-based methods. Categorizing by anomaly determination criterion, the paradigms roughly include the density-estimation, clustering-based and reconstruction-based methods.

As for the density-estimation methods, the classic methods local outlier factor [1] and connectivity outlier factor [10] calculate the local density and local connectivity for outlier determination respectively. DAGMM [16] and MPPCACD [15] integrate the Gaussian Mixture Model to estimate the density of representations.

In clustering-based methods, the anomaly score is always formalized as the distance to the cluster center. SVDD [11] and Deep SVDD [6] gather the representations from normal data to a compact cluster. THOC [7] fuses the multi-scale temporal features from intermediate layers by a hierarchical clustering mechanism and detects the anomalies by the multi-layer distances. ITAD [8] conducts the clustering on decomposed tensors.

The reconstruction-based models attempt to detect the anomalies by the reconstruction error. LSTM-VAE [5] model employed the LSTM backbone for temporal modelling and the Variational AutoEncoder (VAE) for reconstruction. OmniAnomaly [9] further extends the LSTM-VAE model with a normalizing flow and uses the reconstruction probabilities for detection. InterFusion [4] renovates the backbone to a hierarchical VAE to model the inter- and intra-dependency among multiple series simultaneously. GANs [2] are also used for reconstruction-based anomaly detection and perform as an adversarial regularization.

2.2 Transformers for Time Series Analysis

Transformers [13] has shown great power in sequential data processing, such as natural language processing, audio processing and computer vision. For time series analysis, benefiting from the advantage of the self-attention mechanism, Transformers are used to discover reliable long-range temporal dependencies. Unlike the previous usage of Transformers, Anomaly Transformer [14] renovates the self-attention mechanism to the Anomaly-Attention based on the key observation of association discrepancy. AnomalyBERT [3] proposed a data degradation scheme that enables the Transformer-based model to understand the temporal context and has strong capability in detecting real-world anomalies. TranAD [12] uses focus score-based self-conditioning to enable robust multi-modal feature extraction and adversarial training to gain stability. Additionally, model-agnostic meta-learning (MAML) allows us to train the model using limited data.

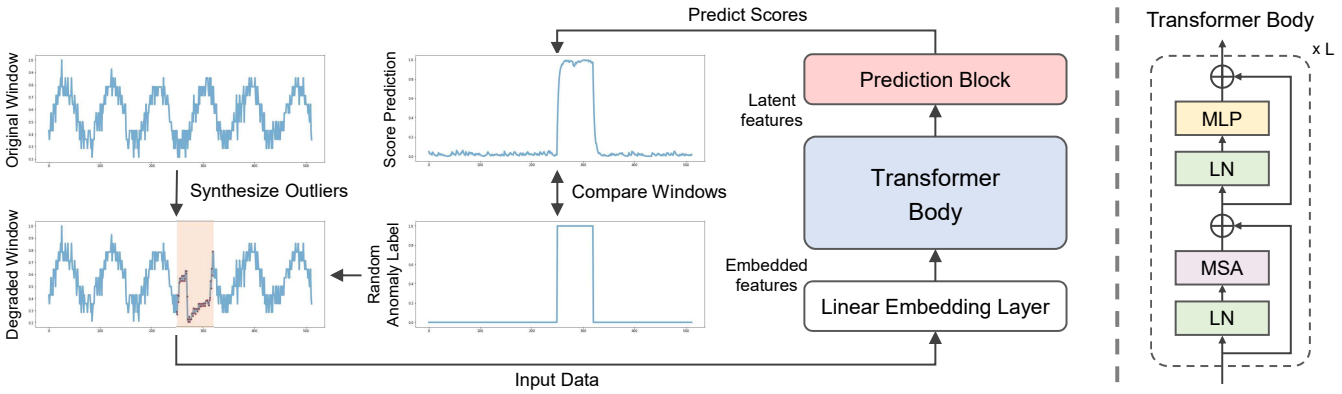

Figure 1: AnomalyBERT architecture. In the training stage, a portion of an input window is randomly replaced and the model is directed to classify the degraded part. The main Transformer body is composed of Transformer layers with 1D relative position bias.

3 MODELS

3.1 AnomalyBERT

AnomalyBERT is a self-supervised transformer model designed for time series anomaly detection using a data degradation scheme. It consists of three main components: a linear embedding layer, a transformer body, and a prediction block [3]. The linear embedding layer projects data patches into embedded features. The transformer body, composed of multiple layers with multi-head self-attention and 1D relative position bias, processes these features to capture temporal context. The prediction block outputs anomaly scores for each data point (Figure 1).

The training objective of AnomalyBERT is based on a binary cross-entropy loss. A portion of the input data is replaced with synthetic outliers, and the model is trained to detect these degraded parts. The loss function is defined as follows:

$$L = -\frac{1}{N}\sum_{t=t_0}^{t_1} 1_{[t_0',t_1']}(t)\log a_t + (1 - 1_{[t_0',t_1']}(t))\log(1-a_t),$$

where N is the window size, $1_{[t_0',t_1']}$ indicates the degraded interval, and a_t is the predicted anomaly score.

3.2 Anomaly Transformer

AnomalyTransformer introduces an association discrepancy mechanism to enhance anomaly detection in time series data. The model architecture includes an encoder-decoder transformer structure with a unique attention mechanism that captures both short-term and long-term dependencies (Figure 2).

The model uses a novel discrepancy loss, which measures the association discrepancy between the normal and abnormal sequences. The objective is to minimize this discrepancy, thereby improving the model's ability to distinguish anomalies from normal data.

3.2.1 Prior-Association. Prior-association captures the inherent structure of normal data by modeling the temporal dependencies within a time series. Let $X = \{x_1, x_2, \ldots, x_T\}$ be a time series. The prior-association for a timestamp t is defined as:

$$P_{i,j} = \frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right),$$

where σ_i is a learnable scale parameter that adapts to the various patterns in the time series, and i, j represent the time points.

3.2.2 Series-Association. Series-association aims to detect anomalies by comparing the learned temporal dependencies with actual observations. It computes the self-attention weights learned from the raw series. For a given timestamp t, the series-association is defined as:

$$S_{i,j} = \text{Softmax}\left(\frac{Q_i K_j^T}{\sqrt{d}}\right),$$

where Q and K are the query and key matrices respectively, and d is the dimension of these matrices.

3.2.3 Minimax Strategy. The Minimax strategy is employed to amplify the difference between normal and abnormal time points by adjusting the association discrepancy. The loss function for input series $X \in \mathbb{R}^{N\times d}$ is formalized as:

$$L_{\text{Total}}(X, \hat{X}, P, S, \lambda) = \|X - \hat{X}\|_F^2 - \lambda\|\text{AssDis}(P, S; X)\|_1,$$

where $\hat{X}$ is the reconstructed input, $\|\cdot\|_F$ is the Frobenius norm, and λ is a trade-off parameter. The association discrepancy AssDis is computed as:

$$\text{AssDis}(P, S; X) = \frac{1}{N}\sum_{i=1}^{N}\text{KL}(P_i\|S_i) + \text{KL}(S_i\|P_i),$$

where KL denotes the Kullback-Leibler divergence.

3.3 TranAD

TranAD is a transformer-based model that uses self-conditioning and adversarial training for anomaly detection in multivariate time series data. The model includes two transformer encoders and two decoders, enabling it to perform robust multi-modal feature extraction and generate reconstructions of the input data (Figure 3).

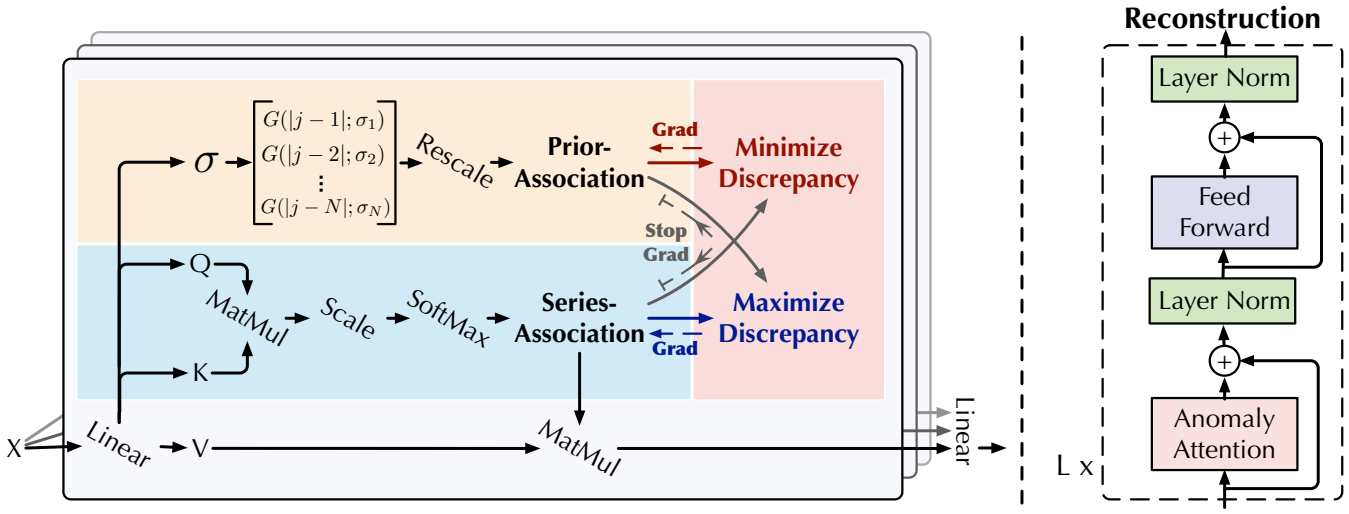

Figure 2: Anomaly Transformer architecture. Anomaly-Attention(left)models the prior-association and series-association simultaneously. In addition to the reconstruction loss, model is optimized by the minimax strategy with a specially-designed stop-gradient mechanism (gray arrows) to constrain the prior- and series- associations for more distinguishable association discrepancy.

TranAD utilizes an adversarial training process, combining reconstruction loss and adversarial loss to enhance anomaly detection. The training process involves two phases: input reconstruction and focus score-based self-conditioning, which guides the model to focus on high-deviation areas.

3.3.1 Self-Conditioning. Self-conditioning is used to enable robust multi-modal feature extraction. It involves conditioning the model on its own outputs to focus on areas with high deviations. Given an input window W, the self-conditioning mechanism updates the window as:

$$W' = W + \alpha \cdot (W - O_1),$$

where O_1 is the initial reconstruction of W and α is a hyperparameter controlling the conditioning strength.

3.3.2 Adversarial Training. Adversarial training involves training the model to distinguish between normal and anomalous data by using a discriminator. The generator G tries to minimize the reconstruction error while the discriminator D maximizes it. The loss functions are defined as:

$$L_G = \|W - G(W)\|_2^2,$$

$$L_D = -\|W - D(G(W))\|_2^2.$$

3.3.3 Model-Agnostic Meta Learning (MAML). MAML is used to optimize the model with limited data. It involves learning an initialization that can quickly adapt to new tasks with few training examples. The meta-learning loss is defined as:

$$L_{\text{MAML}} = \sum_{i=1}^{N} L_i(\theta - \alpha \nabla_\theta L_i(\theta)),$$

where θ are the model parameters, α is the learning rate, and L_i is the loss for task i.

3.3.4 Two-Phase Adversarial Training. TranAD employs a two-phase adversarial training strategy. In the first phase, the model aims to generate an approximate reconstruction of the input window. The reconstruction loss for the first decoder is used as a focus score for the second phase. This process can be summarized as:

$$L_1 = \|O_1 - W\|_2,$$

$$L_2 = \|O_2 - W\|_2.$$

In the second phase, the adversarial training strategy maximizes the difference between the input window and the candidate reconstruction generated by the first decoder:

$$\min_{D_1} \max_{D_2} \|\hat{O}_2 - W\|_2.$$

3.3.5 Evolving Training Objective. The training objective evolves over epochs to balance the reconstruction and adversarial losses. Initially, the reconstruction loss is given higher weight to ensure stable training. As reconstructions improve, the weight shifts towards the adversarial loss:

$$L_1 = \epsilon^{-n}\|O_1 - W\|_2 + (1 - \epsilon^{-n})\|\hat{O}_2 - W\|_2,$$

$$L_2 = \epsilon^{-n}\|O_2 - W\|_2 - (1 - \epsilon^{-n})\|\hat{O}_2 - W\|_2,$$

where n is the training epoch and ϵ is a training parameter close to one.

4 COMPARISON

I reproduced the results in three papers. More precisely, I conducted the comparison experiment with three models on four datasets including MSL, SMAP, SMD and SWaT, and record the precision, recall, f1-score, training time, and inference time. See Table 1 2 3 4 5 6. Reproduction and experiment code is available at: https://github.com/Leon-Huang001208/AnomalyBERT

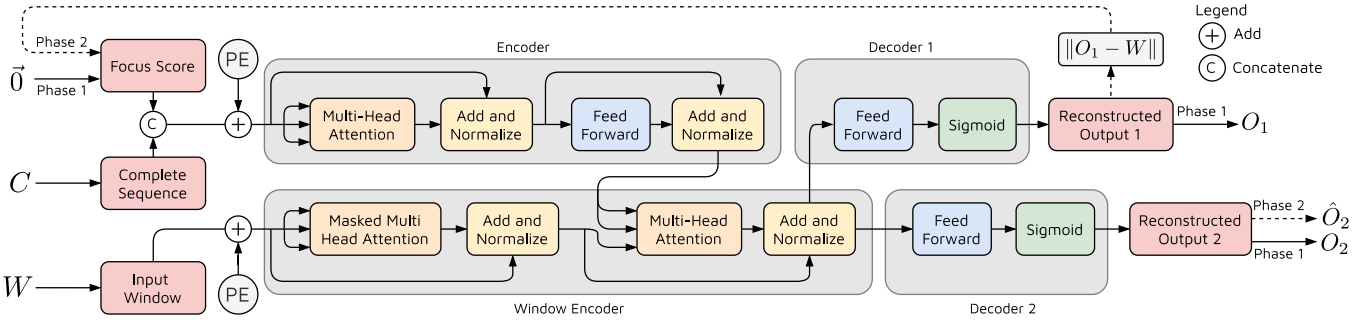

Figure 3: TranAD architecture.

https://github.com/Leon-Huang001208/Anomaly-Transformer
https://github.com/Leon-Huang001208/TranAD

5 CONCLUSION

Transformer-based models have shown significant promise in time series anomaly detection, each offering unique strengths and addressing different challenges. AnomalyBERT excels in handling unlabeled data, AnomalyTransformer is adept at capturing long-term dependencies, and TranAD combines robustness with efficiency. Understanding their differences in model structure, loss functions, advantages, limitations, applicable scenarios, and inference speeds can guide practitioners in selecting the most suitable model for their specific needs.

Dataset	MSL			SMAP			SMD			SWaT		
Metric	P	R	F1	P	R	F1	P	R	F1	P	R	F1
AnomalyBERT	0.561	0.6105	0.585	0.887	0.9435	0.914	0.895	0.773	0.830	0.977	0.878	0.925
AnomalyTran	0.921	0.952	0.936	0.941	0.994	0.967	0.894	0.955	0.923	0.916	0.967	0.941
TranAD	0.904	0.999	0.949	0.804	0.999	0.892	0.926	0.997	0.961	0.976	0.700	0.815

Table 1: Performance comparison of different methods on various datasets.

Method	MSL	SMAP	SMD	SWaT
AnomalyBERT	172	385	69	118
AnomalyTran	31.53	71.82	4.33	10.26
TranAD	0.87	0.758	8.99	0.36

Table 2: Comparison of training times in seconds per epoch.

Method	MSL	SMAP	SMD	SWaT
AnomalyBERT	5.72	2.23	1.96	1.08
AnomalyTran	10.08	21.31	4.36	10.96
TranAD	36.72	17.57	42.33	0.90

Table 3: Comparison of testing times in seconds.

Table 4: Advantages and Limitations of Different Models

Model	**Advantages**	**Limitations**
AnomalyBERT	High detection accuracy, effective with unlabeled data	Complexity in designing synthetic outliers, potential overfitting
AnomalyTransformer	Captures long-term dependencies, robust to anomaly patterns	High computational cost, requires careful hyperparameter tuning
TranAD	Robust to small deviations, efficient with limited data	Increased training complexity, higher computational requirements

Table 5: Applicable Scenarios for Different Models

Model	**Applicable Scenarios**
AnomalyBERT	Industrial monitoring, scenarios with scarce labeled data
AnomalyTransformer	Network traffic analysis, applications requiring high accuracy
TranAD	Rapid detection with minimal data, environments with high data volatility

Table 6: Inference Speed of Different Models

Model	**Inference Speed**
AnomalyBERT	Fast, but computationally intensive due to self-attention
AnomalyTransformer	Moderate, slowed by complex attention mechanism
TranAD	Fast, optimized for parallel processing

REFERENCES

[1] Markus Breunig, Peer Kröger, Raymond Ng, and Joerg Sander. 2000. LOF: Identifying Density-Based Local Outliers. *ACM Sigmod Record* 29, 93–104. https://doi.org/10.1145/342009.335388

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

[3] Yungi Jeong, Eunseok Yang, Jung Hyun Ryu, Imseong Park, and Myungjoo Kang. 2023. AnomalyBERT: Self-Supervised Transformer for Time Series Anomaly Detection using Data Degradation Scheme. arXiv:2305.04468 [cs.LG]

[4] Zhihan Li, Youjian Zhao, Jiaqi Han, Ya Su, Rui Jiao, Xidao Wen, and Dan Pei. 2021. Multivariate Time Series Anomaly Detection and Interpretation using Hierarchical Inter-Metric and Temporal Embedding, See [4], 3220–3230. https://doi.org/10.1145/3447548.3467075

[5] Daehyung Park, Yuuna Hoshi, and Charles C. Kemp. 2017. A Multimodal Anomaly Detector for Robot-Assisted Feeding Using an LSTM-based Variational Autoencoder. arXiv:1711.00614 [cs.RO]

[6] Lukas Ruff, Robert Vandermeulen, Nico Görnitz, Lucas Deecke, Shoaib Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. 2018. Deep One-Class Classification.

[7] Lifeng Shen, Zhuocong Li, and James Kwok. 2020. Timeseries Anomaly Detection using Temporal Hierarchical One-Class Network. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 13016–13026. https://proceedings.neurips.cc/paper_files/paper/2020/file/97e401a02082021fd24957f852e0e475-Paper.pdf

[8] Youjin Shin, Sangyup Lee, Shahroz Tariq, Myeong Shin Lee, OkchulJung, Daewon Chung, and Simon Woo. 2020. Integrative Tensor-based Anomaly Detection System For Satellites. https://openreview.net/forum?id=HJeg46EKPr

[9] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network. 2828–2837. https://doi.org/10.1145/3292500.3330672

[10] Jian Tang, Zhixiang Chen, Ada Fu, and David Cheung. 2002. Enhancing Effectiveness of Outlier Detections for Low Density Patterns. 535–548. https://doi.org/10.1007/3-540-47887-6_53

[11] David Tax and Robert Duin. 2004. Support Vector Data Description. *Machine Learning* 54 (01 2004), 45–66. https://doi.org/10.1023/B:MACH.0000008084.60811.49

[12] Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. 2022. TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data. arXiv:2201.07284 [cs.LG]

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[14] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. 2022. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. arXiv:2110.02642 [cs.LG]

[15] Takehisa Yairi, Naoya Takeishi, Tetsuo Oda, Yuta Nakajima, Naoki Nishimura, and Noboru Takata. 2017. A Data-Driven Health Monitoring Method for Satellite Housekeeping Data Based on Probabilistic Clustering and Dimensionality Reduction. *IEEE Trans. Aerospace Electron. Systems* 53, 3 (2017), 1384–1401. https://doi.org/10.1109/TAES.2017.2671247

[16] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. 2018. Deep Autoencoding Gaussian Mixture Model for Unsupervised Anomaly Detection. In *International Conference on Learning Representations*. https://openreview.net/forum?id=BJJLHbb0-