ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447104

SKIP-STEP CONTRASTIVE PREDICTIVE CODING FOR TIME SERIES ANOMALY DETECTION

Kexin Zhang[1, 2], Qingsong Wen[3,†], Chaoli Zhang[4], Liang Sun[3], Yong Liu[1, 2]

[1] Huzhou Institute of Zhejiang University, Huzhou, China
[2] Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou, China
[3] DAMO Academy, Alibaba Group, Bellevue, WA, USA
[4] School of Computer Science and Technology, Zhejiang Normal University, Jinhua, China

ABSTRACT

Self-supervised learning (SSL) shows impressive performance in many tasks lacking sufficient labels. In this paper, we study SSL in time series anomaly detection (TSAD) by incorporating the characteristics of time series data. Specifically, we build an anomaly detection algorithm consisting of global pattern learning and local association learning. The global pattern learning module builds encoder and decoder to reconstruct the raw time series data to detect global anomalies. To complement the limitation of the global pattern learning that ignores local associations between anomaly points and their adjacent windows, we design a local association learning module, which leverages contrastive predictive coding (CPC) to transform the identification of anomaly points into positive pairs identification. Motivated by the observation that adjusting the distance between the history window and the time point to be detected directly impacts the detection performance in the CPC framework, we further propose a skip-step CPC scheme in the local association learning module which adjusts the distance for better construction of the positive pairs and detection results. The experimental results show that the proposed algorithm achieves superior performance on SMD and PSM datasets in comparison with 12 state-of-the-art algorithms.

Index Terms— Time Series, Anomaly Detection, Self-supervised Learning

1. INTRODUCTION

Time series anomaly detection (TSAD) is one of the challenging tasks in many real-world applications [1, 2, 3, 4, 5], such as server monitoring, process control, influenza detection, etc. Traditional methods usually use feature engineering to generate features and then design feature-based detection algorithms to achieve anomaly detection. As real systems become more and more complex, feature engineering becomes more difficult and traditional methods fail to achieve good detection performance. Deep learning (DL) is a powerful technique for feature learning, which directly extracts features from the data and reduces the efforts in the design of hand-crafted features. Generally, building a successful deep learning model requires a large amount of labeled data. Unfortunately, labeled anomaly data is very limited for time series data, which motivates us to seek a method for anomaly detection using unlabeled data more effectively.

Recently, self-supervised learning (SSL), which allows for learning representations without ground-truth labels, has exerted great power in the fields of computer vision, natural language processing, and signal processing [6]. Contrastive learning (CL) is an important branch of SSL, and it has been applied widely in image classification tasks [7]. The key idea behind CL is to minimize the distance between similar samples and maximize the distance between dissimilar samples, and it has also been adopted in time series analysis [8].

In this paper, we investigate CL for time series anomaly detection. We observe that due to the rarity of anomalies in time series, it is difficult to use anomaly points to construct positive pairs in the CL framework straightforwardly. In other words, the contrastive loss will be large when the constructed positive sample pairs contain anomaly points. Based on this observation, we introduce contrastive predictive coding (CPC) [9] to the TSAD task. The CPC aims to learn representations by predicting the future in latent space through a given history window. The distance between history window and future time points is an important factor affecting the detection results because different distances represent different positive pairs.

Based on the previous analysis, we propose a new TSAD algorithm consisting of two major modules: an autoencoder (AE) global pattern learning module and a skip-step CPC-based local association learning module. The global pattern learning module tries to reconstruct the raw data through an encoder and a decoder, in which the anomaly points have significantly larger reconstruction errors. A limitation of the global module is that it ignores local associations between

†Corresponding author

This work was funded by Collective Intelligence & Collaboration Laboratory (Open Fund Project No. QXZ23012301) and supported by the National Key R&D Program of China under Grant 2021YFB2012300.

anomaly points and their adjacent windows. To address this issue, a local association learning module is designed to leverage CPC to transform the identification of anomaly points into positive pairs identification tasks in CL. The final anomaly score is then calculated by fusing the global anomaly score and local anomaly score, where each time point is determined as normal or abnormal with a threshold. The main contributions are summarized as follows:

- We propose a new framework for TSAD task, which transforms the task of identifying anomalies into the identification of positive and negative samples within the CL framework.
- We design a novel skip-step contrastive scheme that empowers the model to choose appropriate positive samples across varying temporal distances, thus leading to a further enhancement in detection accuracy.
- The proposed method achieves performance comparable or superior to state-of-the-art methods on three popular benchmark datasets.

2. RELATED WORKS

Time series anomaly detection is a very vital and challenging task in practice. Given the time-consuming and labor-intensive nature of acquiring sufficient labeled data, most approaches adopt an unsupervised learning framework. These approaches can be broadly classified into categories, including clustering-based, reconstruction-based, prediction-based, and association-based methods.

The clustering-based methods detect anomalies by measuring the distance between time points and normal pattern cluster centers. THOC [10] uses the fused features from all intermediate layers of dilated recurrent neural network by a differentiable hierarchical clustering mechanism and detects the anomalies by a novel score that measures the distance in the multiple hyperspheres. The reconstruction-based methods typically use an AE architecture to reconstruct the raw data and then detect anomalies based on the reconstruction error [11, 12, 13]. For example, OmniAnomaly [11] captures the normal patterns by learning their robust representations with stochastic variable connection and planar normalizing flow and uses the reconstruction probabilities to determine anomalies. USAD [13] proposes an encoder-decoder architecture within an adversarial training framework that allows it to isolate anomalies while providing fast training. The prediction-based methods employ an autoregressive model to predict the future time points and calculate the prediction error [14, 15]. For example, LNT [15] applies CPC to produce good semantic time series representations and makes predictions of the context at different time horizons. The association-based method is a new anomaly detection approach that learns associations between a time point and its adjacent time points [16, 17]. AnomalyTrans [16] proposes the transformer-based model with a newly designed anomaly-attention mechanism, which can model the prior-association and series-association simultaneously to embody the association discrepancy, and the learned association discrepancy is a criterion for detecting anomalies.

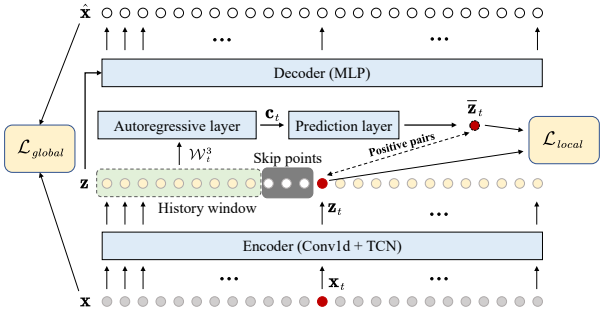

Fig. 1. Overview of the proposed detection framework.

3. METHODOLOGY

Suppose the time series samples has C measurements with length N, and we denote by $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) \in \mathbb{R}^{N \times C}$, where $\mathbf{x}_t \in \mathbb{R}^C$ is the observation at time t. In time series anomaly detection, our goal is to determine whether $\mathbf{x}_t$ is an anomaly point. Fig. 1 shows the structure of the proposed time series anomaly detection algorithm. First, a single 1-D convolutional layer maps the input time point $\mathbf{x}_t$ to embedding space $\mathbf{h}_t = \mathrm{Conv1d}(\mathbf{x}_t)$, where $\mathbf{h}_t \in \mathbb{R}^d$. Next, a temporal convolutional network (TCN) [18] is used to further map $\mathbf{h}_t$ to more informative representation $\mathbf{z}_t = \mathrm{TCN}(\mathbf{h}_t)$. Then, the model is divided into two branches. The first branch uses MLP as a decoder to reconstruct the raw input $\mathbf{x}_t$, i.e., $\hat{\mathbf{x}}_t = \mathrm{MLP}(\mathbf{z}_t)$. This branch learns global information in the data, which makes anomaly points exhibit large reconstruction errors. However, this module ignores the local associations between anomaly points and their adjacent windows, which is very useful for determining whether a time point is normal or abnormal [16]. Therefore, we introduce another new branch, called Skip-Step CPC, to capture the local association between a time point and its adjacent time window. The second branch uses an autoregressive model to summarize the historical window $\mathcal{W}_t$ in the new latent space and produces a context representation $\mathbf{c}_t$. Unlike the original CPC, we found that keeping a certain distance between $\mathcal{W}_t$ and $\mathbf{x}_t$ can better capture local associations.

A simple illustration is shown in Fig. 2. We define $\mathcal{W}_t^d$ to be a historical window with the distance d from $\mathbf{x}_t$, where d represents the number of time points between the last time point of $\mathcal{W}_t$ and $\mathbf{x}_t$. We expect that when $\mathbf{x}_t$ is not an anomaly point, this branch maximizes the mutual information between the $\mathbf{x}_t$ and the context representation of $\mathcal{W}_t^d$. On the contrary, if $\mathbf{x}_t$ is an anomaly point, it is difficult to extract the underlying patterns in common.

Both the AE-based and Skip-Step CPC-based branches are trained to jointly optimize the final loss as

$$\mathcal{L} = \mathcal{L}_{AE} + \mathcal{L}_{CPC}. \tag{1}$$

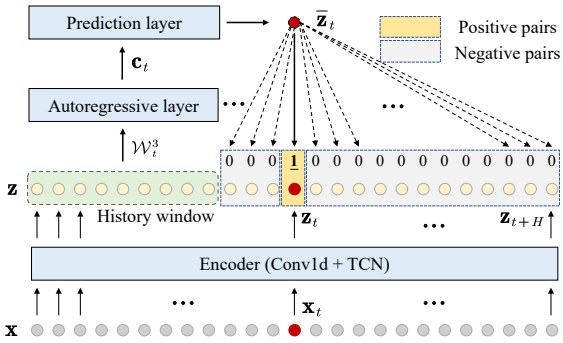

Fig. 2. The strategy of calculating local contrastive scores.

The first term in Eq. (1) is the reconstruction loss, defined as

$$\mathcal{L}_{AE} = \frac{1}{NC}\sum_{i=1}^{N}\sum_{j=1}^{C}(x_i^j - \hat{x}_i^j)^2, \tag{2}$$

where x_i^j is the ground truth and $\hat{x}_i^j$ is the output of the AE-based branch. The second term in Eq. (1) is the InfoNCE loss $\mathcal{L}_{CPC}$, which can maximize a lower bound of mutual information [9], defined as

$$\mathcal{L}_{CPC} = \mathbb{E}_{\mathcal{X}}[-\log \frac{\exp(\mathbf{z}_t^T \mathbf{W}^d \mathbf{c}_t^d)}{\sum_{\mathbf{z}_j \in \mathcal{X}} \exp(\mathbf{z}_j^T \mathbf{W}^d \mathbf{c}_t^d)}], \tag{3}$$

where $\mathcal{X}$ is a set of N random training samples. Each sample contains a history window and a time point. The loss in Eq. (3) is the categorical cross-entropy of classifying the positive sample correctly.

Finally, the output anomaly score for $\mathbf{x}_t$ is calculated by fusing global and local anomaly scores as

$$S(t) = S_{local}(t) * S_{global}(t). \tag{4}$$

The global anomaly score for $\mathbf{x}_t$ is calculated by reconstruction error, defined as

$$S_{global}(t) = (\mathbf{x}_t - \hat{\mathbf{x}}_t)^2. \tag{5}$$

The local anomaly score for $\mathbf{x}_t$ is computed by the cross entropy loss, as shown in Fig. 3. Specifically, we first make a prediction $\bar{\mathbf{z}}_t = \mathbf{W}^d \mathbf{c}_t^d$, then the distribution for the time point is computed based on the similarity between the prediction $\bar{\mathbf{x}}_t$ and representations of all time points in predefined windows $\mathcal{W}_p$. Therefore, the local contrastive score is calculated as

$$S_{local}(t) = -\log \frac{\exp(\mathbf{z}_t^T \bar{\mathbf{z}}_t)}{\sum_{\mathbf{z}_j \in \mathcal{W}_p} \exp(\mathbf{z}_j^T \bar{\mathbf{z}}_t)}. \tag{6}$$

Based on the anomaly score, a predefined threshold δ is used to decide if a point is an anomaly (1) or not (0). The output of $\mathbf{x}_t$ is defined as

$$y(t) = \begin{cases} 1, & S(t) > \delta \\ 0, & S(t) \leq \delta \end{cases}. \tag{7}$$

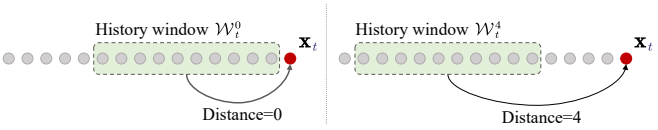

Fig. 3. Illustration of the proposed skip-step CPC scheme. (Constructing positive pairs with different distances)

4. EXPERIMENTS AND DISCUSSION

We evaluate our detection algorithm on three popular benchmark datasets with 12 DL-based baseline models. We evaluate our method on three datasets. (1) SMD (Server Machine Dataset) [11] is a new 5-week-long dataset and is collected from a large Internet company with 38 dimensions. (2) MSL (Curiosity) [14] is a spacecraft anomaly detection dataset and it has 55 dimensions. (3) PSM (Pooled Server Metrics) [19] is collected internally from multiple application server nodes at eBay with 26 dimensions. We use random sampling to obtain training samples, and each sample contains a historical window and a future window. The historical window is with a fixed size of 50 and the future window is set to 20 for all datasets. The time points are labeled as anomalies if their anomaly scores are larger than the predefined threshold δ. The widely-used adjustment strategy [11, 16, 17] that an anomaly segment is considered correctly detected as long as any point in this segment is detected is adopted in this paper. We use precision P, recall R, and F1-score $F1$ as evaluation metrics. F1-score is the harmonic mean of precision and recall. A high value indicates better performance. We compare our model with 12 DL-based methods. The results of the baselines are collected from [16, 17].

4.1. Comparison Results

It can be observed that detection methods that take into account the contextual information of time series tend to yield more favorable results, as seen in the case of AnomalyTrans, which incorporates local associations. In summary, our proposed method achieves better performance than all 11 baseline models on datasets SMD and PSM, and achieves results close to the state-of-the-art on the MSL dataset. The superiority of our model comes from incorporating both reconstruction-based and prediction-based strategies from the global pattern and local context information perspectives to make better decisions.

4.2. Skip-step Scheme Analysis

As aforementioned, adjusting the distance between the history window and the time point to be detected in the proposed skip-step CPC scheme has a direct impact on the detection performance in the proposed method. The underlying reason behind this is that the adjustment of the distance essentially changes the construction of the positive pairs in the CPC framework. To illustrate it, we conduct experiments with different

Table 1. Results of time series anomaly detection on three popular benchmark datasets. The best results are highlighted.

Dataset	SMD			PSM			MSL		
Metric	P	R	F1	P	R	F1	P	R	F1
Deep-SVDD	78.54	79.67	79.10	95.41	86.49	90.73	91.92	76.63	83.58
LSTM	78.55	85.28	81.78	76.93	89.64	82.80	85.45	82.50	83.95
CL-MPPCA	82.36	76.07	79.09	56.02	99.93	71.80	73.71	88.54	80.44
ITAD	86.22	73.71	79.48	72.80	64.02	68.13	69.44	84.09	76.07
LSTM-VAE	75.76	90.08	82.30	73.62	89.92	80.96	85.49	79.94	82.62
BeatGAN	72.90	84.09	78.10	90.30	93.84	92.04	89.75	85.42	87.53
OmniAnomaly	83.68	86.82	85.22	88.39	74.46	80.83	89.02	86.37	87.67
InterFusion	87.02	85.43	86.22	83.61	83.45	83.52	81.28	92.70	86.62
THOC	79.76	90.95	84.99	88.14	90.99	89.54	88.45	90.97	89.69
TS-CP^2	87.42	66.25	75.38	82.67	78.16	80.35	86.45	68.48	76.42
AnomalyTrans	89.40	95.45	92.33	96.91	**98.90**	97.89	92.09	95.15	93.59
DCdetector	83.59	91.10	87.18	97.14	98.74	97.94	**93.69**	**99.69**	**96.60**
Ours	**91.75**	**97.34**	**94.46**	**98.36**	98.74	**98.55**	90.84	94.73	92.75

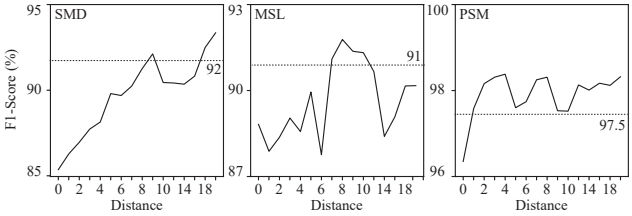

Fig. 4. Anomaly detection results at different distances in the proposed skip-step CPC scheme.

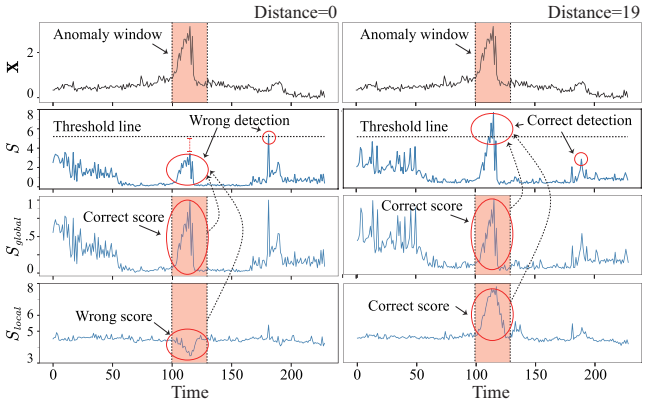

Fig. 5. Case 1 of the distance studies in the proposed skip-step CPC scheme. Left figure: false negative around time index 110 and false positive around time index 180 under distance=0. Right figure: the correct results are obtained under distance=19.

distances on the three benchmark datasets and depict the results in Fig. 4. It can be observed that as the distance increases, the F1-score of SMD also increases, and better detection performance is achieved when the distance is greater than 8. The best detection performance for the MSL is obtained when the distance is 8. The F1-score of PSM increases first and then remains stable, and it is least sensitive to distance. The results in Fig. 4 show that we need to set different distances to construct reasonable positive pairs under different datasets for better anomaly detection results.

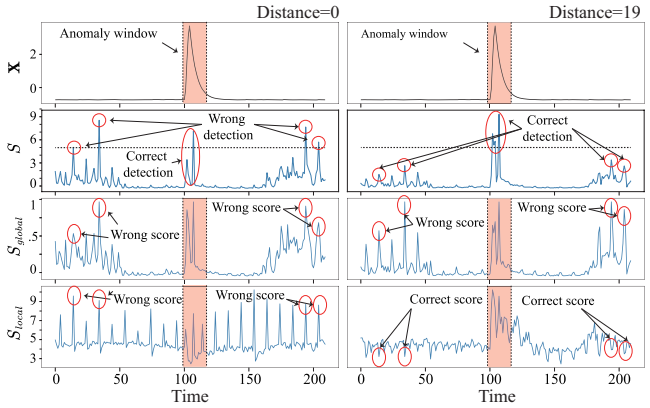

Fig. 6. Case 2 of the distance studies in the proposed skip-step CPC scheme. Left figure: false positive around time indices 15, 35, 190, and 210 under distance=0. Right figure: the correct results are obtained under distance=19.

4.3. Case Analysis

We take the most distance-sensitive dataset as an example, i.e., SMD. The distances are set to 0 and 19, respectively. The first is the false negative and false positive detection case as shown in Fig. 5. We observe that local contrastive scores exhibit different patterns at anomaly locations when different distances of positive samples are given. Although the reconstruction error score is correct when the distance is 0, the wrong local score leads to false negative detection. On the contrary, the local score is correct when the distance is 19, which means the model learns the correct pattern. If we only use the global reconstruction score, it would result in a false positive detection (a spike with right circles in Fig. 5). Therefore, the correct local contrastive score and correct global reconstruction score amplify the final anomaly score, making the anomalous time points easier to detect. The second is a case that contains both true detection and false positive detection, as shown in Fig. 6. In this case, we observed that although the local detection module provides the correct score at distance 0, it also leads to excessively large scores (spikes with right circles) elsewhere, detecting some normal points as anomaly points. However, the local anomaly score only shows a larger value at the location of the anomaly when the distance is 19.

5. CONCLUSION AND FUTURE WORK

This paper investigates the application of SSL for TSAD and we propose a new detection method that combines the AE-based global pattern learning module and the skip-step CPC-based local association learning module. The proposed method achieves superior performance to the state-of-the-art results on two datasets. In future, we plan to extend the current research, including the adaptive construction of positive and negative samples and the introduction of adversarial training to enhance the robustness of the detection model.

6. REFERENCES

[1] Qingsong Wen, Linxiao Yang, Tian Zhou, and Liang Sun, "Robust time series analysis and applications: An industrial perspective," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2022, pp. 4836–4837.

[2] Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A Lozano, "A review on outlier/anomaly detection in time series data," *ACM Computing Surveys (CSUR)*, vol. 54, no. 3, pp. 1–33, 2021.

[3] Jingkun Gao, Xiaomin Song, Qingsong Wen, Pichao Wang, Liang Sun, and Huan Xu, "RobustTAD: Robust time series anomaly detection via decomposition and convolutional neural networks," *KDD Workshop MileTS*, 2020.

[4] Andrew A Cook, Göksel Mısırlı, and Zhong Fan, "Anomaly detection for iot time-series data: A survey," *IEEE Internet of Things Journal*, vol. 7, no. 7, pp. 6481–6494, 2019.

[5] Longyuan Li, Junchi Yan, Qingsong Wen, Yaohui Jin, and Xiaokang Yang, "Learning robust deep state space for unsupervised anomaly detection in contaminated time-series," *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2022.

[6] Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales, "Self-supervised representation learning: Introduction, advances, and challenges," *IEEE Signal Processing Magazine*, vol. 39, no. 3, pp. 42–62, 2022.

[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020.

[8] Kexin Zhang, Qingsong Wen, Chaoli Zhang, Rongyao Cai, Ming Jin, Yong Liu, James Zhang, Yuxuan Liang, Guansong Pang, Dongjin Song, and Shirui Pan, "Self-supervised learning for time series analysis: Taxonomy, progress, and prospects," *arXiv preprint arXiv:2306.10125*, 2023.

[9] Aäron van den Oord, Yazhe Li, and Oriol Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[10] Lifeng Shen, Zhuocong Li, and James Kwok, "Time-series anomaly detection using temporal hierarchical one-class network," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020, vol. 33, pp. 13016–13026.

[11] Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei 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*, 2019, p. 2828–2837.

[12] Lifeng Shen, Zhongzhong Yu, Qianli Ma, and James T. Kwok, "Time series anomaly detection with multiresolution ensemble decoding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 9567–9575.

[13] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga, "Usad: Unsupervised anomaly detection on multivariate time series," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, p. 3395–3404.

[14] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom 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, p. 387–395.

[15] Tim Schneider, Chen Qiu, Marius Kloft, Decky Aspandi-Latif, Steffen Staab, Stephan Mandt, and Maja Rudolph, "Detecting anomalies within time series using local neural transformations," *CoRR*, vol. abs/2202.03944, 2022.

[16] Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," in *International Conference on Learning Representations (ICLR)*, 2022.

[17] Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang 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*, New York, NY, USA, 2023, KDD '23, p. 3033–3045, Association for Computing Machinery.

[18] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *CoRR*, vol. abs/1803.01271, 2018.

[19] 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*, 2021, p. 2485–2494.