WHEN WILL IT FAIL?: ANOMALY TO PROMPT FOR FORECASTING FUTURE ANOMALIES IN TIME SERIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Recently, time series forecasting, which predicts future signals, and time series anomaly detection, which identifies abnormal signals in given data, have achieved impressive success. However, in the real world, merely forecasting future signals or detecting anomalies in existing signals is not sufficiently informative to prevent potential system breakdowns, which lead to huge costs and require intensive human labor. In this work, we tackle a challenging and under-explored problem of time series anomaly prediction. In this scenario, the models are required to forecast the upcoming signals while considering anomaly points to detect them. To resolve this challenging task, we propose a simple yet effective framework, Anomaly to Prompt (A2P), which is jointly trained via the forecasting and anomaly detection objectives while sharing the feature extractor for better representation. On top of that, A2P leverages Anomaly-Aware Forecasting (AAF), which derives the anomaly probability by random anomaly injection to forecast abnormal time points. Furthermore, we propose Synthetic Anomaly Prompting (SAP) for more robust anomaly detection by enhancing the diversity of abnormal input signals for training anomaly detection model. As a result, our model achieves state-of-the-art performances on seven real-world datasets, proving the effectiveness of our proposed framework A2P for a new time series anomaly prediction task.

1 INTRODUCTION

Recently, time series analyses have been broadly explored because they are crucial in real-world applications. Major tasks in time series analysis can be divided into three folds: time series forecasting (Zhang & Yan, 2022; Shen et al., 2020; Zhou et al., 2023; Liu et al., 2024; Nie et al., 2023; Wang et al., 2022a; Zhou et al., 2022; 2021), time series anomaly detection (Audibert et al., 2020; Xu et al., 2022; Su et al., 2019a; Yang et al., 2023), and time series classification (Lu et al., 2022; Early et al., 2023; Xiao et al., 2022). Among them, time series forecasting is essential for predicting future trends and patterns, aiding in effective decision-making across various fields. Time series anomaly detection is also critical for identifying unusual patterns or events within data, enabling timely intervention and mitigation of potential risks or issues. Both time series forecasting and anomaly detection are important in terms of practicality, e.g., risk mitigation, proactive management, and cost savings of a system.

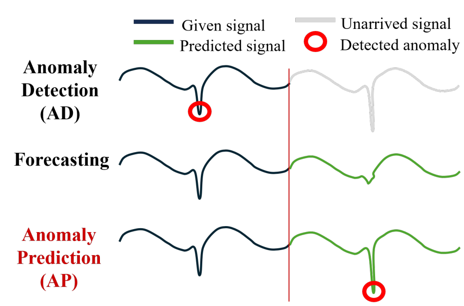

Figure 1: Comparison among different scenarios of existing time series anomaly detection, forecasting, and a newly proposed anomaly prediction.

However, merely forecasting future signals or detecting anomalies in already elapsed time signals has limited applicability in crucial real-world scenarios. For example, medical doctors need to predict potential abnormal accidents in patients' biometric data to make decisions about their health in advance. However, existing forecasting or anomaly detection methods alone give limited informative messages to the doctors in this scenario. With only predicted future signals, the doctor cannot use it for decision making directly unless the doctor scrutinizes the signal, consuming considerable time and effort. With only the result of anomaly detection of signals from already elapsed time, the doctor has no choice but to guess about the status of future. Another example case can be the maintenance of industrial systems, where a prediction of future abnormal events is crucial because companies or users can minimize costs from abrupt system failure. Despite the significance of the above scenarios, how to predict when abnormal events will happen in the future is under-explored. Moreover, we aim to predict at what time points the anomalies can happen, which is much more challenging than simply detecting anomalies from the given signal.

In this paper, to address the above issue, we first propose a new scenario called **Anomaly Prediction** (**AP**), which forecasts and detects the anomaly points in the future signal. To fulfill the proposed AP task, the model needs to foresight *the time steps on which abnormal events are possible to happen* for a given signal. Existing anomaly detection methods detect abnormal time points from the given signal, and forecasting methods merely predict how future signals will look like. In contrast, in AP, the model should be able to detect anomalies in the predicted signal as shown in Figure 2. Most of the existing state-of-the-art time series anomaly detection approaches have a limitation in that they adopt a point adjustment technique proposed in Audibert et al. (2020) which only assesses the detection of anomaly segments, not the accurate abnormal time points. Thus, they cannot appropriately evaluate whether the model locates the positions of anomalies correctly or not. However, in the proposed AP, it is crucial to predict the exact time steps of anomalies as possible in order to save costs for maintenance systems of industrial devices or medical treatments, etc., where proactive management is necessary.

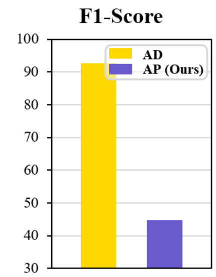

Figure 2: Comparison of F1-scores for existing time series anomaly detection task (AD) and our proposed Anomaly Prediction task (AP) in MBA dataset.

The comparison between the average performances of the existing anomaly detection (AD) scenario which detects anomalies from already arrived signals, and our proposed AP scenario, where the model detects anomalies from predicted future signals is conducted in Figure 2. As shown in Figure 2, we empirically found that a naïve combination of the time series forecasting model and time series anomaly detection model fails at predicting future anomalies, where the anomaly detection model detects anomalies from predicted signals that are the outputs of the forecasting model. The reason for the failure is quite intuitive: existing forecasting models are trained on only normal signals and predict them, thereby overlooking the prominence of abnormality in abnormal time points when predicting future signals. As a result, anomaly detection models fail at detecting anomalies because the forecasting models rather reduce the degree of abnormality of anomaly time points, which makes it difficult to detect them for anomaly detection models.

To effectively resolve the novel scenario, we propose a simple yet effective framework, **Anomaly to Prompt (A2P)**, which is composed of **A**nomaly-**A**ware **F**orecasting (**AAF**) and **S**ynthetic **A**nomaly **P**rompting (**SAP**). AAF aims to consider the existence of anomalies in real world into the training process of forecasting. To achieve this, we utilize Anomaly-Aware Forecasting Network (AAFN) which is pre-trained before the main training to learn the probabilities of being an anomaly in a signal. Along with the method to enhance the capability to forecast anomalous time steps, we introduce a novel Synthetic Anomaly Prompting method with Anomaly Prompt Pool (APP) to improve the robustness of anomaly detection so that it can cope with more diverse signals. Anomaly prompts, which are learnable parameters, are utilized to intensify the diversity of signals used for reconstruction in the anomaly detection model. They are based on signal-adaptive prompt tuning to guide signals in mimicking abnormal features, leveraging an anomaly prompt pool that contains instructions for transforming normal signals into anomalous ones. Furthermore, we adopt a shared backbone architecture that can learn a unified representation for performing forecasting and anomaly detection at once.

Our main contributions can be summarized as follows:

- For the first time, we propose a new scenario called Anomaly Prediction (**AP**), which aims to point out at what time points the abnormal events are likely to occur in the future given the arrived signals. This is more challenging than existing time series anomaly detection since existing forecasting and anomaly detection models are only trained to grasp features of normal signals. To tackle AP effectively, we propose a new framework, coined Anomaly to Prompt (A2P).

- We propose an unprecedented method for forecasting time points with anomalies. To achieve this, we introduce an Anomaly-Aware Forecasting (AAF) mechanism to enhance the forecasting ability of future signals containing anomalies.
- We propose a novel Synthetic Anomaly Prompting (SAP) method to simulate anomalies. Moreover, we introduce two novel loss objectives, Intra-Signal Anomaly Discrepancy loss and Inter-Signal Anomaly Divergence loss, to learn anomaly prompt pool. These objectives enable our model to effectively capture the characteristics of anomalies from the training data via pre-training of anomaly prompts.
- We conducted comprehensive experiments on various real-world datasets to show the effectiveness of our proposed methods and demonstrate that our method outperforms the state-of-the-art methods.

2 Related Work

Time Series Forecasting. Time series anomaly detection, which is the task of forecasting future signals based on historical observations, is important in terms of practicality, such as network monitoring, weather forecasting, economics and finance, and electricity forecasting. Previous works on time series have achieved strong prediction performance by leveraging advances in sequence modeling machine learning methods and deep neural networks such as RNN (Hochreiter & Schmidhuber, 1997; Tokgöz & Ünal, 2018; Abdel-Nasser & Mahmoud, 2019), GNN (Jiang & Luo, 2022; Wang et al., 2022b; Panagopoulos et al., 2021), and CNN (Bai et al., 2018; Livieris et al., 2020) to capture temporal dependencies. Recently, Transformer (Vaswani et al., 2017) have begun to be actively used for time series forecasting since it has become very prominent in natural language processing (Zhou et al., 2021; 2022; Wu et al., 2021; Liu et al., 2021; Cirstea et al., 2022; Zhang & Yan, 2022; Nie et al., 2023; Zhou et al., 2023). However, existing forecasting models are trained with only normal signals, excluding the possibility of abnormal events in the signals. Therefore, even though they can effectively learn the features of normal signals, they fail to predict anomalous signals appropriately. Moreover, the forecasting model alone cannot give useful information for any decision-making directly because a human being should investigate the inherited meanings of predicted signals to make use of them, which is time-consuming and labor intensive. Therefore, in this paper, we address this issue by introducing a new task that a model should forecast from given signals and detect anomalies in them.

Time Series Anomaly Detection. Multivariate time series anomaly detection is a crucial problem for many applications and has been widely studied. Most of the previous studies are mainly performed in an unsupervised manner considering the restriction on access to abnormal data. Traditional approaches include the density-estimation and clustering-based methods. However, since they cannot consider the complex temporal dynamics most of the recent works focus on deep learning-based approaches. Among them, reconstruction-based approaches (Shen et al., 2021; 2020; Li et al., 2019; Su et al., 2019c; Zhou et al., 2019; Yang et al., 2023) find latent representations of normal time series data for reconstruction. Recently, Xu et al. (2022) proposes a new association-based method, which applies the learnable Gaussian kernel to introduce the adjacent-concentration bias for better reconstruction. Another recent reconstruction-based model DCdetector (Yang et al., 2023) achieves a similar goal in a much more general and concise way with a dual-attention self-supervised contrastive-type structure. The existing anomaly detection models focus on detecting the time points of anomalies in already elapsed time, which has limited application in real-world. In this paper, we address a more practical and challenging scenario, Anomaly Prediction to predict the anomalies in future signals which have not arrived yet.

Synthetic Anomalies for Time Series. The intuitive approach to overcome the absence of abnormal data at training time is to artificially generate them during training time. However, synthetic anomalies and their associated data augmentation techniques have not been widely studied in time series anomaly detection field. Zhang et al. (2022) proposes a data augmentation module to increase the robustness of the network by generating various sequence anomalies. However, it does not consider the specific abnormal time points of augmented signal at training time although it is crucial to distinguish specific abnormal time points at test time. Another approach for synthetic anomaly is introduced in Carmona et al. (2021), which injects synthetic abnormal points into a portion of normal signal. Then it trains the contrastive classifier to create embeddings where the part of the signal without injected anomalies is distant from the entire signal with anomalies. However, it operates only at the embedding level, making it difficult to precisely identify the exact abnormal points, whereas ours considers both specific anomaly time points and their embedding outputs. Recently, Goswami et al. (2023) injects 9 types of anomalies to select the most accurate model for a given dataset without labels. However, all the methods above can only simulate limited forms of anomalies without considering the inherent features of the input signal. In this paper, we design a simple yet effective Synthetic Anomaly Pool method to infuse signal-adaptive learnable anomalies, considering the abnormality of specific time points. Moreover, we propose an energy score-based loss function, Intra-Signal Anomaly Discrepancy loss, and Inter-Signal Anomaly Divergence loss for imitation to enable the synthesized signal to effectively imitate abnormal characteristics.

3 Method

In this section, we introduce a new scenario called Anomaly Prediction, to foresight potential anomalies in future signals. We first define our proposed task in detail in section 3.1. Then, we explain the architecture of A2P, a unified shared backbone network to perform both forecasting and anomaly detection at once in section 3.2. To tackle our challenging scenario effectively, in section 3.3, we introduce a new approach called Anomaly-Aware Forecasting (AAF) for more precise forecasting of abnormal time points. Furthermore, we propose a novel method coined Synthetic Anomaly Prompting (SAP) which trains a newly proposed Anomaly Prompt Pool (APP) from randomly injected anomaly in section 3.4. Finally, we summarize the total objective function in section 3.5.

3.1 Scenario Description: Time Series Anomaly Prediction

Time series anomaly prediction is a novel scenario that aims to pinpoint the exact time steps of anomaly points in the upcoming signals. Specifically, for a given input signal $X_{in} \in \mathbb{R}^{L_{in} \times C}$, the final goal is to obtain the binary results of anomaly detection $O \in \mathbb{R}^{L_{out}}$ from the predicted signal $\hat{X}_{out} \in \mathbb{R}^{L_{out} \times C}$, where L_{in} and L_{out} are the lengths of the input and predicted signals, respectively, and C is the number of channels in the signal. To perform anomaly prediction, we need a network for time series forecasting denoted as Θ_F, and a network for time series anomaly detection denoted as Θ_{AD}. Therefore, O can be written as $O = \Theta_{AD} \circ \Theta_F(X_{in}) = \Theta_{AD}(\hat{X}_{out})$, where $\hat{X}_{out} = \Theta_F(X_{in})$. For the evaluation of anomaly prediction performance, the F1-score is used as existing time series anomaly detection methods do. The difference in the measurement from the existing methods is that the Point Adjustment (PA) proposed in Audibert et al. (2020) cannot be adopted in original way, since in anomaly prediction, it is important to identify specific time points rather than to identify the existence of anomaly segments. Therefore, we alleviate PA for our metric, which is explained in section 4.

3.2 Unified Architecture for Anomaly Prediction

Existing time series forecasting models such as Nie et al. (2023); Wang et al. (2022a); Zhou et al. (2023); Wu et al. (2021); Zhou et al. (2021; 2022); Liu et al. (2021) and anomaly detection models like Xu et al. (2022); Yang et al. (2023) consider the train set comprised of only normal signals. Accordingly, although the target task of each model is different, both networks for time series forecasting and anomaly detection attempt to capture the characteristics of normal time series data in common. Inspired by this point, we adopt a shared backbone to establish a unified architecture to learn the representations of normal signals for both the forecasting and anomaly detection models, as shown in Figure 4.

Specifically, in our framework, several base layers of transformer blocks denoted as θ are shared, while other specific parts, the embedding layers (e_F and e_{AD}) and output layers (o_F and o_{AD}) to construct Θ_F and Θ_{AD}, exist separately, *i.e.*, $\Theta_F = \{e_F, o_F, \theta\}$ and $\Theta_{AD} = \{e_{AD}, o_{AD}, \theta\}$. By sharing the backbone network, our model can accumulate general knowledge for both time series forecasting and anomaly detection effectively, resulting in rich representations and performance improvements. We analyze the effectiveness of our unified framework in Section 4.3 with details.

3.3 Anomaly-Aware Forecasting

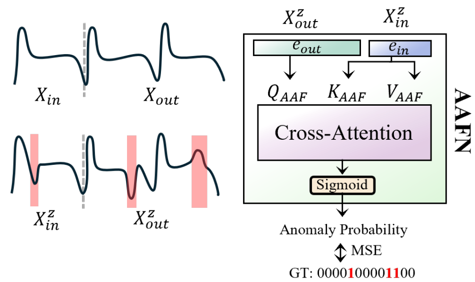

Figure 3: Training scheme of Anomaly-Aware Forecasting Network (AAFN). Input signal X_{in} and its corresponding future signal X_{out} are randomly injected to be in charge of abnormal signal since labeled anomaly is scarce to be used in train time in the real world. Then, each of the signals is fed into cross-attention in AAFN, followed by an activation function.

In this work, we propose a novel method called Anomaly-Aware Forecasting (AAF), which enhances the accuracy of future signal prediction by explicitly accounting for anomalies in prior signals. Unlike traditional forecasting models that treat all past data equally, our approach incorporates an additional module to improve robustness in dynamic and unpredictable environments. The core of this method is the Anomaly-Aware Forecasting Network (AAFN), which is pre-trained to learn the complex relationships between prior signal anomalies and future trends. This pre-training step enables the network to anticipate how past anomalies might influence upcoming signals, thereby providing a more informed and accurate forecast during the main training phase.

The main purpose of AAF is to learn the relation between abnormal features inherent in a prior signal and its following future signal. To this end, we exploit AAFN which is composed of embedding layers, an attention layer, and an activation layer as shown in figure 3. The inputs of AFFN are X_{out}^z and X_{in}^z, which are the results of random anomaly injection among seasonal, global, trend, contextual, shapelet anomaly type from X_{out} and X_{in}, respectively, following the algorithm used in (Darban et al., 2025). The query for attention in ATTN is $e_{out}(X_{out}^z)$, which is the target that we want to know about, while key and value are $e_{in}(X_{in}^z)$, the ground for assessing the abnormality of X_{out}^z when e_{out} and e_{in} refers to the embedding layers for X_{out}^z and x_{in}^z. The output of AAFN is trained to indicate the probability of being anomaly for each time step. This output is compared to ground truth label y_{out}^z of X_{out}^z, and mean squared error is used for the loss term. As a result, the final loss term for training AAFN in advance of the main training is as follows:

$$\mathcal{L}_{AAFN} = MSE(\sigma(Attn(e_{out}(X_{out}^z), e_{in}(X_{in}^z))), y_{out}^z), \tag{1}$$

where σ is the activation function, sigmoid function, and $Attn$ is the cross attention layer.

3.4 Synthetic Anomaly Prompting

Signal-Adaptive Anomaly Prompting. To accurately predict anomaly points, forecasting the future signal from a prior signal while considering the existence of anomaly points is crucial. To tackle this challenge, we propose a novel approach, named Synthetic Anomaly Prompting (**SAP**), which utilizes synthetic anomaly prompts for our model to predict future abnormal signals effectively. For SAP, we integrate a new Anomaly Prompt Pool (**APP**) into our unified architecture, as shown in Figure 4. The purpose of APP, which is a set of additional trainable parameters P, is to guide an input signal to behave like an abnormal signal, by infusing the anomaly prompts in the pool into the original signal.

In detail, APP is defined as $P = \{(k_1, Z_1), (k_2, Z_2), \cdots, (k_M, Z_M)\}$, where $Z_m \in \mathbb{R}^{L_z \times D}$ and $k_m \in \mathbb{R}^D$ denote the m-th anomaly prompt and its corresponding key, respectively, L_z and D are the token length of single anomaly prompt and the embedding dimension, and M is the number of anomaly prompts. Moreover, to select the number of N best-matched prompts with the input signal X_{in} in the pool, we introduce a feature extractor $f_{ftr}(\cdot)$, which is a simple three-layer transformer architecture with a [CLS] token, as a query function, *i.e.*, $q(X_{in}) = f_{ftr}([CLS; X_{in}])$.

The process of our proposed anomaly synthesis method using APP is displayed in figure 4. First, we pre-train the feature extractor $f_{ftr}(\cdot)$ with the train set, which will be used to select the most relevant anomaly prompt from APP. After the training of $f_{ftr}(\cdot)$, it is frozen and used only for the retrieval of features from normal signals in stage 1 and stage 2. Second, we train Anomaly Prompt Pool with input data which is converted into anomaly, using naive random anomaly injection. We injected anomaly in the same way that was used in Darban et al. (2025), as well. The input signal passes through the feature extractor to obtain the query $q(X_{in})$. This query is then matched against the keys in the Anomaly Prompt Pool, and the prompts corresponding to the top-N closest keys are attached to the embedded input $\widetilde{X}_{in} \in \mathbb{R}^{L_{in} \times D}$, where $\widetilde{X}_{in} = e_{AD}(X_{in})$. Note that the synthesis of anomaly is executed at the embedding level, not the raw input level, which enables more diverse prompting in

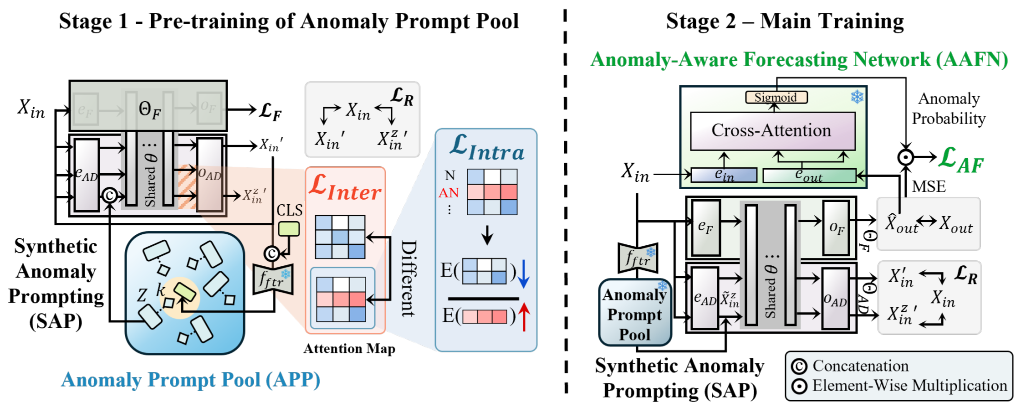

Figure 4: Overview of model architecture and training strategies. Our model first pre-trains Anomaly Prompt Pool (APP) in the first stage, by injecting random anomaly to train data. After pre-training APP before the main training phase, it is frozen in the main training phase since it already holds the knowledge of real-world anomalies. Otherwise, when there are no abnormal time steps in train set, the first stage is skipped, and the model trains APP jointly with Θ_F and Θ_{AD}. Note that abnormal signals are only used in the first stage.

high dimensions. Finally, the simulated embedding of anomaly $\widetilde{X}^z_{in}$ is defined as follows:

$$\widetilde{X}^z_{in} = [Z_{s_1}; \cdots ; Z_{s_N}; \widetilde{X}_{in}], \quad s_i \in \mathbf{S}, \tag{2}$$

$$\mathbf{S} = \underset{\{s_i\}_{i=1}^{N} \subseteq [1,M]}{\operatorname{argmax}} \sum_{i=1}^{N} \gamma\left(q(X_{in}), k_{s_i}\right), \tag{3}$$

where the score function γ is cosine similarity, which is for calculating how each anomaly prompt is related to each normal signal. $[\cdot;\cdot]$ denotes the concatenation. The selected prompt tokens are attached to the input tokens of $\widetilde{X}_{in}$, after passing through the embedding layer e_{AD} before the first layer of θ, to transform the original normal signal into an abnormal signal. The output anomaly prompts are then removed before the final projection of each o_F and o_{AD}, to match the dimension of each of them.

Intra-Signal Anomaly Discrepancy Loss for Synthetic Anomaly Prompting. To effectively train the proposed APP, we design a new loss function called Intra-Signal Anomaly Discrepancy loss ($\mathcal{L}_{Intra}$). The role of Intra-Signal Anomaly Discrepancy loss is to *spur the anomaly prompts in APP to catalyze $\widetilde{X}_{in}$ to behave like anomalies*. Based on the fact that abnormal points tend to be more associated with adjacent time points (Xu et al., 2022), we assume that the distribution of the attention map of a plausible abnormal point is uneven whereas that of normal signal is distributed evenly. Encouraged by this assumption, we introduce the energy score (Liu et al., 2020) to force the distribution of attention map at abnormal points in the initial layer to have a lower score compared to that of the normal signal.

Subsequently, we regulate the energy scores within that signal, such that the portions corresponding to normal points have reduced energy scores (becoming more uniform), while the portions corresponding to abnormal points have increased energy scores (becoming more uneven), as depicted in Figure 4. Here, we randomly select portions within the entire signal to become abnormal points when training with only normal signals in main training phase. Along with a term to make abnormal signals, we add an additional term to pull the selected keys closer to the corresponding features of normal signals as follows:

$$\mathcal{L}_{Intra} = \frac{E\left(A^l_N\left(\widetilde{X}^z_{in}\right)\right)}{E\left(A^l_{AN}\left(\widetilde{X}^z_{in}\right)\right)} - \lambda_k\, \gamma\left(f_{ftr}(X_{in}), k_m\right), \tag{4}$$

where λ_k is a scalar to weight the loss and E is the energy scoring function A^l_{AN} and A^l_N refer to the portions of the output attention map of the last layer $\boldsymbol{A}^l$ where l is the number of layers, from input embedding w, which correspond to abnormal points and normal points within the synthetic abnormal signal, respectively.

In addition, we employ an additional loss term that aims to reconstruct $\widetilde{X}_{in}^{z}$ back to its original normal form X, along with the existing reconstruction loss for anomaly detection as follows:

$$\mathcal{L}_R = \frac{1}{2}\left(\left\|X_{in} - X_{in}^{z}{}'\right\|^2 + \|X_{in} - X'_{in}\|^2\right), \tag{5}$$

where $X_{in}^{z}{}' = o_{AD}(\theta(\widetilde{X}_{in}^{z}))$ is the reconstruction output of synthesized abnormal input embedding and $X'_{in} = \Theta_{AD}(X_{in})$ is that of original normal input signal. Note that the SAP with the Intra-Signal Anomaly Discrepancy loss can be used even when there is no anomaly available in train set, since the SAP works in a self-supervised way. Along with $\mathcal{L}_R$, we adopt the existing forecasting loss $\mathcal{L}_F$ for forecasting the future signals, defined as follows:

$$\mathcal{L}_F = \left\|\hat{X}_{in} - X_{in}\right\|^2. \tag{6}$$

In main training stage, a pre-trained Anomaly-Aware Forecasting Network as explained in section 3.3 is used to output anomaly probability. Therefore, in main training, the final loss term regarding forecasting is as follows:

$$\mathcal{L}_{AF} = \left\|\hat{X}_{in} - X_{in}\right\|^2 \odot AAFN(X_{in}, \hat{X}_{out}), \tag{7}$$

where $\hat{X}_{in}$ is $\Theta_F(X_{in})$, $AAFN$ is Anomaly-Aware Forecasting Network, and $\odot$ is element-wise multiplication. By considering the errors in anomaly time steps more than other time steps, the network can be trained to focus on abnormal areas.

Inter-Signal Anomaly Divergence Loss for Pre-training of Anomaly Prompt Pool. To make the model be prepared for detecting more diverse anomalies, we pre-train the Anomaly Prompt Pool which holds the knowledge of characteristics of anomalies, using only abnormal signals injected randomly while training jointly with Θ_F and Θ_{AD}. The pre-trained Anomaly Prompt Pool can then be used to infuse plausible anomalies into the model in the later main training phase. For the efficient pre-training of Anomaly Prompt Pool, we introduce a novel Inter-Signal Anomaly Divergence loss ($\mathcal{L}_{Inter}$) to guide the anomaly prompts in the Anomaly Prompt Pool to prompt the signals to imitate actual abnormal signals as follows:

$$\mathcal{L}_{Inter} = -KL(A_l(\widetilde{X}_{in}'^{z}), A_l(\widetilde{X}'_{in})). \tag{8}$$

Here, we obtain the reconstruction output X'_{in} which plays a role of pseudo-normal signal corresponding to anomaly signal X_{in}. Then, the model attaches anomaly prompts from Anomaly Prompt Pool to pseudo-normal embedding $\widetilde{X}'_{in}$ to simulate anomaly, which results in $\widetilde{X}_{in}'^{z}$. Since we aim to train APP to add abnormalities into the signal, the features of synthetic anomaly and pseudo-normal input features are trained to be distinct with $\mathcal{L}_{Inter}$, which serves to intensify the gap between the features of pseudo-normal signal and synthetic anomaly signal.

3.5 Total Objective Function

The total objective function of our proposed framework is summarized as follows:

$$\mathcal{L}_{Total} = \begin{cases} \mathcal{L}_{AAFN}, & \text{for pre-training of AAFN,} \\ \lambda_R\,\mathcal{L}_R + \lambda_F\,\mathcal{L}_F + \lambda_{Intra}\,\mathcal{L}_{Intra} + \lambda_{Inter}\,\mathcal{L}_{Inter}, & \text{for pre-training of APP,} \\ \lambda_R\,\mathcal{L}_R + \lambda_{AF}\,\mathcal{L}_{AF}, & \text{for main training.} \end{cases} \tag{9}$$

The objectives for the pre-training of APP and main training phases are the same except for the Inter-Signal Anomaly Divergence loss and Intra-Signal Anomaly Discrepancy loss, where $\lambda = \{\lambda_R, \lambda_F, \lambda_{Intra}, \lambda_{Inter}\}$ is a set of coefficients for weighting each loss term. We used all four coefficients of 1 as default values for experiments. Note that only abnormal signals are used for training during APT, whereas only normal signals are used in the main training phase. After the pre-training phase, the proposed APP is frozen during the main training phase.

4 Experiments

4.1 Experimental Setup

Dataset Configurations. We evaluated our method on six real-world multivariate datasets and one real-world univariate dataset: Both 1) MSL (Mars Science Laboratory rover) and 2) SMAP (Soil

Moisture Active Passive satellite) (Hundman et al., 2018) were published by NASA with 55 and 25 dimensions, respectively, which contained the telemetry anomaly data derived from the Incident Surprise Anomaly (ISA) reports of spacecraft monitoring systems. 3) PSM (Pooled Server Metrics) (Abdulaal et al., 2021) was collected internally from multiple application server nodes at eBay with 25 dimensions. 4) SMD (Server Machine Dataset) (Su et al., 2019b) was a 5-week-long dataset that was collected from a large Internet company with 38 dimensions. 5) SWaT (Secure Water Treatment) (Mathur & Tippenhauer, 2016) was obtained from 51 sensors of the critical infrastructure system under continuous operations. 6) WADI (Water Distribution) (Ahmed et al., 2017) was an extension of SWaT and a distribution system comprising a larger number of water distribution pipelines with 123 dimensions. 7) MBA (MIT-BIH Supraventricular Arrhythmia Database) (Moody & Mark, 2001) is a set of electrocardiogram recordings from four patients, composed of two distinct types of irregularities (supraventricular contractions or premature heartbeats). The datasets were divided into a training set, and a test set. Abnormal data exists only in the test set.

Baselines and Evaluation Metrics. We compared our model with various combinations of existing forecasting models and anomaly detection models, considering them as our baselines. For forecasting models, we adopted state-of-the-art models, PatchTST (Nie et al., 2023), MICN (Wang et al., 2022a), GPT2 (Zhou et al., 2023), and iTransformer (Liu et al., 2024). Regarding anomaly detection models, we adopted reconstruction-based methods, AnomalyTransformer (Xu et al., 2022) and DCDetector (Yang et al., 2023). We used F1-score (F1) as the main evaluation metric. If not mentioned, the scores reported in the tables indicate F1-scores. In addition, F1-score was calculated without point adjustment introduced in Audibert et al. (2020). Instead, we used F1-score with tolerance t, which denotes the time steps to tolerate the error in anomaly detection. For example, when a model predicts that a time step i is anomaly, the real ground-truth anomaly time points from $[i-t, i+t]$ are considered to be correctly detected before the calculation of F1-Score. We also validated the results on various tolerance t on MSL dataset in Table 8.

Hyperparameters. The hyperparameters used are mentioned in Appendix A, and the sensitivity results on various parameter values are shown in Appendix C.

4.2 Anomaly Prediction Results

The results of the Anomaly Prediction experiments are demonstrated in Tables 1 and 2. For the F1-score, our model consistently outperforms the baselines, showing the effectiveness of our proposed Anomaly-Aware Forecasting and Synthetic Anomaly Prompting.

Note that ours were effective in datasets from various domains, which implies that our methods are robust to the various statistics of datasets. Especially, even when the length of output signal was longer, the gaps between comparing methods and ours became large. It indicates that our methods are robustly advantageous in more challenging scenarios where it is essential to examine the status of the distant future to prevent possible hazards in the future. Furthermore, we evaluated the effectiveness of our proposed A2P on univariate time series, the MBA dataset which is a bio-medical data, in Table 2. It demonstrates that A2P can be generalized to the univariate case as well.

Model		L_{out}			Avg. F1
F	AD	100	200	400	
P-TST	AT	49.10±0.03	43.94±0.09	31.09±0.08	41.38±0.07
	DC	47.17±0.06	49.17±0.03	<u>42.82±0.02</u>	46.39±0.04
MICN	AT	50.18±0.06	<u>49.75±0.02</u>	42.80±0.02	47.58±0.03
	DC	<u>53.73±0.04</u>	48.46±0.07	41.33±0.05	<u>47.84±0.05</u>
GPT2	AT	41.85±0.05	46.07±0.06	33.64±0.02	40.52±0.04
	DC	52.98±0.06	47.86±0.04	41.94±0.06	47.59±0.05
iTransformer	AT	47.54±0.06	48.27±0.07	39.37±0.02	45.06±0.50
	DC	51.21±0.04	45.43±0.10	37.80±0.06	44.81±0.07
A2P (Ours)		**58.66±0.01**	**50.59±0.07**	**48.11±0.14**	**52.45±0.07**

Table 2: Anomaly Prediction results on univariate MBA dataset. The best results are in bold and the second bests are underlined. All experiments were conducted with 3 random seeds.

L_{out}	Model		Dataset						Avg. F1
	F	AD	MSL	PSM	SMAP	SWAT	SMD	WADI	
100	P-TST	AT	41.91±1.28	14.51±1.03	15.11±0.61	12.19±1.02	33.84±0.39	58.02±0.04	29.26±0.73
		DC	25.23±3.49	10.45±4.85	9.66±0.47	11.36±4.50	15.77±4.39	18.50±13.99	15.16±5.28
	MICN	AT	42.61±0.01	13.17±0.00	15.23±0.10	10.97±0.33	29.88±1.09	58.88±0.04	28.46±0.26
		DC	42.40±0.68	14.04±0.41	11.33±4.20	11.41±2.24	30.97±0.39	24.14±9.48	22.38±2.90
	GPT2	AT	41.39±3.82	13.33±0.00	15.32±0.57	11.25±0.28	32.66±1.03	54.26±0.04	28.03±0.96
		DC	4.56±6.45	4.75±6.72	5.00±3.58	9.37±0.07	11.17±13.89	16.35±2.89	8.53±5.60
	iTransformer	AT	42.19±0.34	14.06±0.74	15.19±0.40	11.25±0.25	32.71±0.96	59.20±0.05	29.10±0.46
		DC	40.86±1.85	12.77±1.09	6.66±3.22	13.99±8.12	7.81±10.43	17.32±1.08	16.57±4.30
	A2P (Ours)		**45.04±0.00**	**14.77±0.00**	**16.06±0.00**	**15.00±0.01**	**35.73±0.00**	**61.71±0.02**	**31.39±0.01**
200	P-TST	AT	38.20±0.32	14.79±0.41	14.89±0.24	11.48±0.57	32.79±0.74	54.49±0.04	27.77±0.39
		DC	24.43±5.40	5.70±5.49	8.81±1.54	9.76±5.05	14.35±9.75	24.75±5.38	14.63±5.43
	MICN	AT	42.64±0.01	13.41±0.00	15.59±0.45	11.41±0.34	32.38±1.41	49.79±0.01	27.54±0.37
		DC	41.05±0.55	2.38±3.37	14.88±1.17	10.26±6.36	31.88±2.64	48.79±3.69	24.87±2.96
	GPT2	AT	42.25±0.54	14.69±0.82	14.76±1.03	10.93±0.79	33.62±0.85	52.13±0.02	28.06±0.67
		DC	7.44±10.52	8.44±6.16	6.56±4.96	11.70±0.00	15.48±14.01	26.52±14.00	12.69±8.28
	iTransformer	AT	42.95±2.27	14.03±0.55	15.15±0.44	11.58±0.47	33.70±0.46	51.34±0.04	28.12±0.71
		DC	38.46±3.42	10.47±2.59	7.25±4.79	12.84±4.94	8.69±9.21	25.70±10.25	17.23±5.87
	A2P (Ours)		**48.75±0.06**	**20.69±0.02**	**20.28±0.01**	**15.85±0.00**	**38.15±0.01**	**58.70±0.08**	**33.74±0.03**
400	P-TST	AT	40.53±0.87	14.74±0.25	15.03±0.28	11.20±0.26	34.05±0.72	44.76±0.05	26.72±0.40
		DC	26.23±3.95	9.83±4.72	7.50±0.93	6.11±2.13	23.28±0.20	19.76±8.22	15.45±3.36
	MICN	AT	41.06±0.01	14.83±0.00	15.51±0.22	11.21±0.21	30.45±1.03	49.79±0.01	7.14±0.25
		DC	41.29±1.05	8.97±6.34	11.93±1.66	17.86±1.64	18.84±13.37	27.40±4.31	21.05±4.73
	GPT2	AT	42.19±1.87	14.60±0.00	14.20±0.50	11.24±0.36	32.51±1.52	48.08±0.01	27.14±0.71
		DC	22.13±11.67	11.02±2.56	9.58±4.16	9.78±0.06	18.13±3.80	28.92±17.93	16.59±6.70
	iTransformer	AT	41.64±1.78	14.68±0.22	14.91±0.39	11.06±0.22	33.54±2.03	45.05±0.05	26.81±0.78
		DC	37.04±5.55	9.79±5.80	6.53±1.81	16.30±0.27	11.99±16.32	31.05±17.03	18.78±7.80
	A2P (Ours)		**50.16±0.11**	**30.95±0.02**	**27.08±0.03**	**24.61±0.01**	**53.43±0.08**	**51.66±0.08**	**39.65±0.06**

Table 1: Anomaly Prediction results on multivariate cases. The best results are in bold and the second bests are underlined. All experiments were conducted with 3 random seeds and the average values were reported when $L_{in} = 100$.

4.3 Analysis

Ablation Study. To further examine the effectiveness of our novel methods, we thoroughly conducted ablation studies. The ablation results of Intra-Signal Anomaly Discrepancy loss $\mathcal{L}_{Intra}$ and Inter-Signal Anomaly Divergence loss $\mathcal{L}_{Inter}$ within Signal Adaptive Prompting are indicated in Table 3 to see the impact of each loss term. As shown in Table 3, 1) guiding the energy of attention map regarding abnormal time points to be higher compared to normal time points and 2) driving the synthetic anomaly feature to be distinct from pseudo-normal signals improved Anomaly Prediction performances, respectively.

In addition, Table 4 demonstrates the ablation of Anomaly-Aware Forecasting. The use of AAF enhanced the performance of Anomaly Prediction, indicating that considering abnormal time points for forecasting was effective in detecting future anomalies. In order to investigate the impact of sharing transformer layers between the forecasting model and the anomaly detection model, we conducted ablation experiments regarding the effectiveness of the shared backbone as demonstrated in Table 5.

Sharing the layers of backbone for forecasting and anomaly detection remarkably enhanced the performances, implying that sharing the knowledge of forecasting and anomaly detection helped to enrich the representation learning of time series signals.

We evaluated the impact of pre-training of Anomaly Prompt Pool as shown in Table 6. When the pre-training stage of APP is removed, there was a significant performance drop, highlighting the effectiveness of pre-training APP so that APP can be utilized to synthesize abnormal features in main training stage.

$\mathcal{L}_{Intra}$	$\mathcal{L}_{Inter}$	MBA	MSL	WADI	Avg. F1
✗	✗	35.52	40.70	55.35	43.81
✓	✗	47.31	42.10	58.11	49.17
✗	✓	53.26	44.70	59.70	52.55
✓	✓	**58.66**	**45.04**	**61.71**	**55.14**

Table 3: Ablation results for the proposed loss terms $\mathcal{L}_{Intra}$ and $\mathcal{L}_{Inter}$ in SAP, when $L_{in} = L_{out} = 100$.

AAF	MBA	MSL	WADI	Avg. F1
✗	54.95	44.72	60.37	53.35
✓	**58.66**	**45.04**	**61.71**	**55.14**

Table 4: Ablation for the proposed Anomaly-Aware Forecasting (AAF).

Shared	MBA	MSL	WADI	Avg. F1
✗	44.73	44.23	58.46	49.14
✓	**58.66**	**45.04**	**61.71**	**55.14**

Table 5: Ablation for the shared transformer backbone.

Pre-T	MBA	MSL	WADI	Avg. F1
✗	38.37	41.36	57.32	45.68
✓	**58.66**	**45.04**	**61.71**	**55.14**

Table 6: Ablation for the pre-training of APP.

Results on Forecasting. As shown in Table 7, our proposed A2P was advantageous at predicting more accurate future signals. The performance improvements in not only Anomaly Prediction but also forecasting indicate that our proposed approaches contributed to learning the representations of both normal and abnormal signals effectively, compared to the baseline. Notably, when L_{out} was significantly longer, our proposed A2P outperformed at forecasting future signals with anomalies, indicating the capability of handling long-term signals.

L_{out}	100		200		400	
Metric	MSE	MAE	MSE	MAE	MSE	MAE
P-TST + AT	6.632	0.314	6.665	0.304	8.489	0.328
A2P (Ours)	**2.142**	**0.081**	**2.035**	**0.078**	**2.649**	**0.086**

Table 7: The forecasting performances of the baseline (the combination of PatchTST and AnomalyTransformer) and A2P (Ours) in MSL dataset.

Results on Diverse Tolerances. In the Anomaly Prediction task, it is crucial to detect the exact time steps of anomalies. In this regard, the general Point Adjustment strategy is not fit to the Anomaly Prediction task. Therefore, we define t as the number of time steps to allow errors in anomaly detection outputs before and after each time step, which is used for controlling the difficulty of the task. We conducted experiments by varying t from 10 to ∞, where ∞ is equivalent to the existing point adjustment setting. As shown in Table 8, our proposed A2P outperformed on all t, implying that A2P can be used in diverse scenarios from situations where strict localization of time step is required, to more relaxed scenarios. In our experiments, we used $t = 50$ as our default setting.

t	10	20	50	100	∞	Avg. F1
P-TST+AT	12.95	22.75	41.91	57.38	93.41	45.68
A2P (Ours)	**13.50**	**23.87**	**45.04**	**60.94**	**94.96**	**47.66**

Table 8: Anomaly Prediction results on various tolerances t on MSL dataset.

5 Conclusion

In this paper, we first addressed a novel scenario, named Anomaly Prediction (AP), where the model needs to detect abnormal time points from unarrived future signals. We tackled this issue by establishing a unified architecture that shares the feature extractor for forecasting and anomaly detection models. In addition, we employed to use synthetic anomalies in train time, whereas traditional time series forecasting and anomaly detection models were trained with only normal time series features, limiting their generalizability to abnormal signals. We proposed two effective approaches, Anomaly-Aware Forecasting (AAF) and Synthetic Anomaly Prompting (SAP). In AAF, we designed an Anomaly-Aware Forecasting Network to help the model forecast time steps with anomalies. In SAP, we defined an anomaly prompt pool which learns how to prompt the input signals to have anomalous features. In addition, we devised two novel loss terms, energy loss and Inter-Signal Anomaly Divergence loss, to make the anomaly-prompted features more anomalous. We achieved state-of-the-art performances on the Anomaly Prediction task in seven real-world datasets, demonstrating the effectiveness of our methods through comprehensive experiments. We hope our pioneering attempt to predict future anomalies provides an opportunity to anticipate potential breakdowns, while also opening up a new direction for research.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

References

Mohamed Abdel-Nasser and Karar Mahmoud. Accurate photovoltaic power forecasting models using deep lstm-rnn. *Neural computing and applications*, 31:2727–2740, 2019.

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*, pp. 2485–2494, 2021.

Chuadhry Mujeeb Ahmed, Venkata Reddy Palleti, and Aditya P Mathur. Wadi: a water distribution testbed for research in the design of secure cyber physical systems. In *Proceedings of the 3rd international workshop on cyber-physical systems for smart water networks*, pp. 25–28, 2017.

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*, pp. 3395–3404, 2020.

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*, 2018.

Chris U Carmona, François-Xavier Aubet, Valentin Flunkert, and Jan Gasthaus. Neural contextual anomaly detection for time series. In *IJCAI*, 2021.

Razvan-Gabriel Cirstea, Chenjuan Guo, Bin Yang, Tung Kieu, Xuanyi Dong, and Shirui Pan. Triformer: Triangular, variable-specific attentions for long sequence multivariate time series forecasting–full version. *IJCAI*, 2022.

Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu C Aggarwal, and Mahsa Salehi. Carla: Self-supervised contrastive representation learning for time series anomaly detection. *Pattern Recognition*, 157:110874, 2025.

Joseph Early, Gavin KC Cheung, Kurt Cutajar, Hanting Xie, Jas Kandola, and Niall Twomey. Inherently interpretable time series classification via multiple instance learning. *arXiv preprint arXiv:2311.10049*, 2023.

Mononito Goswami, Cristian Ignacio Challu, Laurent Callot, Lenon Minorics, and Andrey Kan. Unsupervised model selection for time series anomaly detection. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=gOZ_pKANaPW`.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

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*, pp. 387–395, 2018.

Weiwei Jiang and Jiayun Luo. Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications*, 207:117921, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

Dan Li, Dacheng Chen, Baihong Jin, Lei Shi, Jonathan Goh, and See-Kiong Ng. Mad-gan: Multivariate anomaly detection for time series data with generative adversarial networks. In *International conference on artificial neural networks*, pp. 703–716. Springer, 2019.

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International conference on learning representations*, 2021.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33:21464–21475, 2020.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. *ICLR*, 2024.

Ioannis E Livieris, Emmanuel Pintelas, and Panagiotis Pintelas. A cnn–lstm model for gold price time-series forecasting. *Neural computing and applications*, 32:17351–17360, 2020.

Wang Lu, Jindong Wang, Xinwei Sun, Yiqiang Chen, and Xing Xie. Out-of-distribution representation learning for time series classification. *arXiv preprint arXiv:2209.07027*, 2022.

Aditya P Mathur and Nils Ole Tippenhauer. Swat: A water treatment testbed for research and training on ics security. In *2016 international workshop on cyber-physical systems for smart water networks (CySWater)*, pp. 31–36. IEEE, 2016.

George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine*, 20(3):45–50, 2001.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *ICLR*, 2023.

George Panagopoulos, Giannis Nikolentzos, and Michalis Vazirgiannis. Transfer graph neural networks for pandemic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4838–4845, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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

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*, volume 35, pp. 9567–9575, 2021.

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*, pp. 2828–2837, 2019a.

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*, pp. 2828–2837, 2019b.

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*, pp. 2828–2837, 2019c.

Alper Tokgöz and Gözde Ünal. A rnn based time series approach for forecasting turkish electricity load. In *2018 26th Signal processing and communications applications conference (SIU)*, pp. 1–4. IEEE, 2018.

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

Huiqiang Wang, Jian Peng, Feihu Huang, Jince Wang, Junhui Chen, and Yifei Xiao. Micn: Multi-scale local and global context modeling for long-term series forecasting. In *The Eleventh International Conference on Learning Representations*, 2022a.

Lijing Wang, Aniruddha Adiga, Jiangzhuo Chen, Adam Sadilek, Srinivasan Venkatramanan, and Madhav Marathe. Causalgnn: Causal-based graph neural networks for spatio-temporal epidemic forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 12191–12199, 2022b.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Qiao Xiao, Boqian Wu, Yu Zhang, Shiwei Liu, Mykola Pechenizkiy, Elena Mocanu, and Decebal Constantin Mocanu. Dynamic sparse network for time series classification: Learning what to "see". *Advances in Neural Information Processing Systems*, 35:16849–16862, 2022.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. *ICLR*, 2022.

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*, pp. 3033–3045, 2023.

Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. Tfad: A decomposition time series anomaly detection architecture with time-frequency analysis. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pp. 2497–2507, 2022.

Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *The eleventh international conference on learning representations*, 2022.

Bin Zhou, Shenghua Liu, Bryan Hooi, Xueqi Cheng, and Jing Ye. Beatgan: Anomalous rhythm detection using adversarially generated time series. In *IJCAI*, volume 2019, pp. 4433–4439, 2019.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 11106–11115, 2021.

Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *International conference on machine learning*, pp. 27268–27286. PMLR, 2022.

Tian Zhou, Peisong Niu, Liang Sun, Rong Jin, et al. One fits all: Power general time series analysis by pretrained lm. *Advances in neural information processing systems*, 36:43322–43355, 2023.