

Graph Mixture of Experts and Memory-augmented Routers for Multivariate Time Series Anomaly Detection

Xiaoyu Huang^{1,2}, Weidong Chen^{*1}, Bo Hu¹, Zhendong Mao¹

¹University of Science and Technology of China, Hefei, China

²Institute of Plasma Physics, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, China
hhxy@mail.ustc.edu.cn, {chenweidong, hubo, zdmao}@ustc.edu.cn

Abstract

Multivariate time series (MTS) anomaly detection is a critical task that involves identifying abnormal patterns or events in data that consist of multiple interrelated time series. In order to better model the complex interdependence between entities and the various inherent characteristics of each entity, the graph neural network (GNN) based methods are widely adopted by existing methods. In each layer of GNN, node features aggregate information from their neighboring nodes to update their information. In doing so, from shallow layer to deep layer in GNN, original individual node features continue to be weakened and more structural information, *i.e.*, from short-distance neighborhood to long-distance neighborhood, continues to be enhanced. However, research to date has largely ignored the understanding of how hierarchical graph information is represented and their characteristics that can benefit anomaly detection. Existing methods simply leverage the output from the last layer of GNN for anomaly estimation while neglecting the essential information contained in the intermediate GNN layers. To address such limitations, in this paper, we propose a Graph Mixture of Experts (Graph-MoE) network for multivariate time series anomaly detection, which incorporates the mixture of experts (MoE) module to adaptively represent and integrate hierarchical multi-layer graph information into entity representations. It is worth noting that our Graph-MoE can be integrated into any GNN-based MTS anomaly detection method in a plug-and-play manner. In addition, the memory-augmented routers are proposed in this paper to capture the correlation temporal information in terms of the global historical features of MTS to adaptively weigh the obtained entity representations to achieve successful anomaly estimation. Extensive experiments on five challenging datasets prove the superiority of our approach and each proposed module.

Code — <https://github.com/dearlexie1128/Graph-MoE>

Introduction

Detecting anomalies in multivariate time-series data is crucial for ensuring security and preventing financial loss in industrial applications, where devices like servers and engines are monitored using multivariate time-series entities, organizations need efficient systems to identify potential issues

quickly (Ren et al. 2019). However, anomalies are typically rare in both real-world application scenarios and existing public datasets, making data labeling both challenging and costly. Thus, we focus on unsupervised multivariate time series anomaly detection in this paper.

An effective unsupervised strategy is modeling the dataset into a distribution. As machine learning advances (Guo et al. 2024; Gao, Chen, and Xu 2023; Gao, Zhang, and Xu 2020; Hu et al. 2023; Xu et al. 2024b,c,a; Ma et al. 2024, 2021, 2022), many existing methods (Rasul et al. 2021b,a) have been explored alongside this strategy to achieve density estimation of the distribution. However, a key challenge with this approach is that many existing methods ignore the interdependencies between time series, which are critical for accurately modeling the data distribution and performing efficient density estimates for anomaly detection. Graph neural network (GNN) based methods have emerged as a promising solution for enhancing anomaly detection in time-series data by modeling the dependencies between entities. Deng and Hooi (2021) propose to combine structure learning with GNNs to detect anomalies in MTS. MTAD-GAT (Zhao et al. 2020) utilizes graph attention networks to model spatial and temporal dependencies. GRLeN (Zhang, Zhang, and Tsung 2022) learns a probabilistic relation graph for multivariate time series and utilizes hidden variables to capture spatial dependencies. Cai et al. (2021) present a method based on graph neural networks for anomaly detection in dynamic graphs.

Although previous GNN-based methods have achieved promising results, they simply leverage the output from the last layer of GNNs to model the relations between entities. However, as stated in previous analysis works (Chen, Wang, and Li 2021; Chen and Wang 2022; Xu et al. 2018; He et al. 2024; Wang et al. 2024a; Tian, Chen, and Song 2021; Yang et al. 2024b,a) of GNNs, a common limitation of existing method in MTS anomaly detection is that GNNs are essentially “homogeneous” across the whole graph, *i.e.*, forcing all nodes to share the same aggregation mechanism, regardless of the differences in their node features or neighborhoods. That might be suboptimal in graph-based relationship modeling, *e.g.*, when some nodes may require information aggregated over longer ranges while others prefer shorter-range local information. As illustrated in the upper example in Fig. 1, the average number of neighborhood nodes

^{*}Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

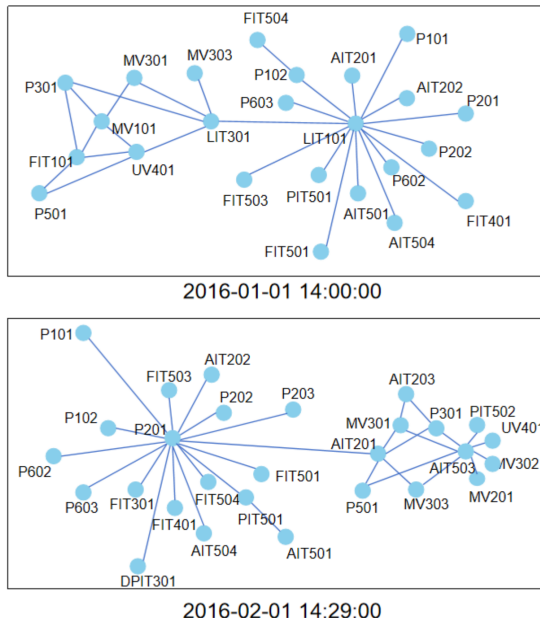


Figure 1: Graph structure of two real-data in MTGFlow (Zhou et al. 2023). Each node indicates an entity of multivariate time-series signals of a certain time step.

within 1-hop, 2-hop, and 3-hop are 2.3, 12.5, and 18.9, respectively, which shows that the network needs to go deep into multi-layer GNNs to establish more global relations. A single-layer GNN can only establish the simplest local relations. Thus, the intermediate layers of GNNs capture information from different multi-hop neighborhoods, *i.e.*, from short-distance to long-distance neighborhoods, which is vital for comprehensively understanding the complex relations and dependencies in MTS data.

In this paper, we propose a Graph Mixture of Experts (Graph-MoE) network, an unsupervised method for multivariate time series (MTS) anomaly detection, to address the aforementioned challenges. Unlike previous GNN-based methods that solely leverage the output from the last layer of GNN to model the interdependences between entities, our Graph-MoE comprehensively utilizes all intermediate information of multi-layer GNNs, that is, it considers the hierarchical node (entity) representations in domains of variable-distance neighborhoods. Specifically, the proposed Graph-MoE incorporates the mixture of experts (MoE) framework to adaptively represent and integrate hierarchical information of different GNN layers into entity representations. In each GNN layer, a specific expert network is implemented with the global-guided entity attention block for the intra-level aggregation of node representations. Moreover, we propose a memory-augmented router, which uses a memory unit to hold the global historical features of the MTS to mine inter-series correlation patterns between time series of the current state, predicting the importance coefficients of experts and facilitating the inter-level aggregation of entity representations. It is worth noting that our Graph-MoE can be integrated into any GNN-based MTS anomaly detection

method in a plug-and-play manner. Extensive experiments on five challenging datasets demonstrate the superiority of our approach and each proposed module. In short, our main contributions are summarized as follows:

- In this paper, we propose a novel Graph Mixture of Experts (Graph-MoE) network for MTS anomaly detection, which extends the expert model at each specific layer of the GNNs. By bringing in experts at different GNN layers with each focusing on different patterns, the Graph-MoE model is able to more fully capture the complex dependencies between entities and the various inherent characteristics of each entity in MTS data.
- We propose a memory-augmented router module to adaptively integrate the outputs of multi-level GNN layers as well as model the relationship between entities better. Our memory-augmented router module leverages the memory mechanism to store the patterns of global historical temporal features to mine the inter-level time-series correlation, achieving effective inter-level aggregation.
- Extensive experiments on five public benchmarks (*e.g.*, SWaT, WADI, PSM, MSL, and SMD) demonstrate the effectiveness of our method and each proposed module, *e.g.*, reaching 87.2 and 94.2 on AUROC%, respectively, on SWaT and WADI datasets.

Related Work

Time Series Anomaly Detection

Time series anomaly detection refers to the identification and discovery of abnormal patterns or behaviors in time series data. However, due to the complexity and volume of time series data, obtaining a well-labeled dataset is challenging (Choi et al. 2021), unsupervised methods are often used to detect temporal anomalies. A typical method is to use auto-encoder (AE) (Kingma and Welling 2013). However, the robustness and generalization capabilities of AEs are limited, which led to the development of VAEs (Su et al. 2019). Additionally, there are graph-based methods in unsupervised learning algorithms. MTGNN (Wu et al. 2020) learns and models relationships between time series variables using a Graph Convolutional Network (GCN). Graph WaveNet (Wu et al. 2019) models the spatio-temporal graph structure through graph convolution to detect anomalies. GANF (Dai and Chen 2022) and MTGFlow (Zhou et al. 2023) leverage graph structures combined with normalizing flow techniques to perform density estimation for anomaly detection in series.

Mixture of Experts

The Mixture of Experts (MoE) is originally developed as an ensemble method that assigns input data to different “expert” submodels, each processing a specific part or feature of the input. To adapt MoE for multi-task learning, Ma et al. (2018) introduce weight-sharing among multiple tasks. Shazeer et al. (2017) introduce the Sparsely-Gated MoE, consisting of thousands of feed-forward sub-networks. Further improvements in the stability and efficiency of sparse MoE models are made by (Lewis et al. 2021). Lepikhin et al.

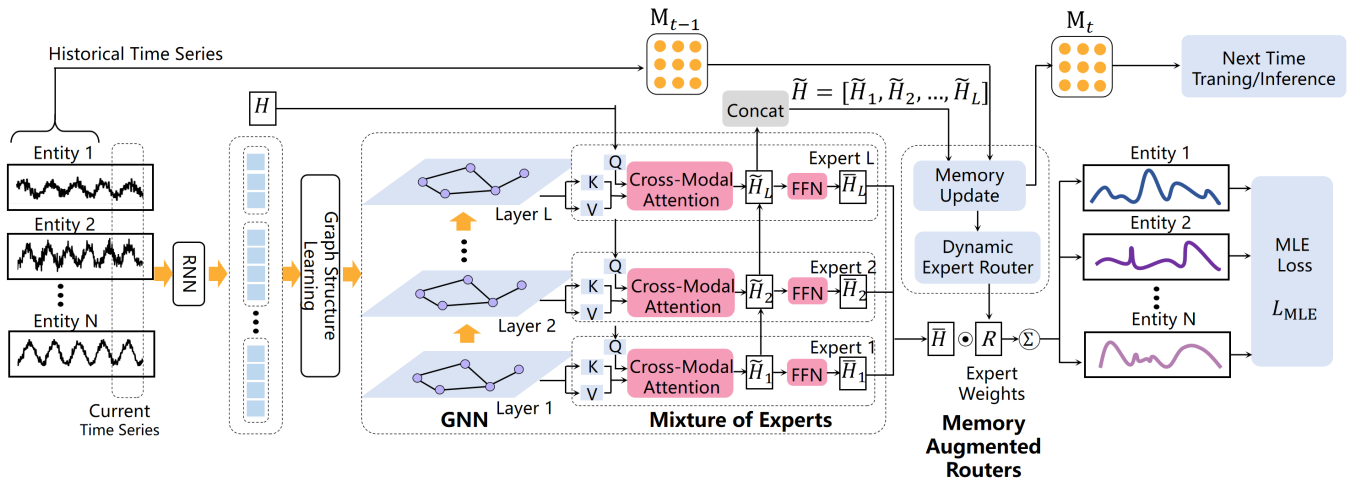


Figure 2: The overview of our proposed Graph-MoE network. It mainly consists of 1) graph mixture of experts and 2) memory-augmented routers.

(2021) utilize MoE to create V-MoE, a vision transformer with sparse activation. Liu et al. (2024b) reveal the Mixture of Multimodal Experts (MoME), enhancing Multimodal Relation Extraction (MRE) through a mixture of hierarchical visual context learners.

Memory-augmented Networks

Growing interest has been attracted to memory-augmented models for improving long-term dependency modeling. LongMem (Wang et al. 2024b) enhances language models with a memory network to store and retrieve long-term context for better language modeling. In recommendation systems, MA-GNN (Ma et al. 2020) uses a graph neural network for short-term context and a shared memory network for long-term dependencies, while DMAN (Tan et al. 2021) segments behavior sequences and uses dynamic memory blocks to combine short-term and long-term user interests for more accurate recommendations. These models highlight the importance of memory mechanisms.

Methodology

Preliminaries

We present a brief introduction to our baseline in this section to provide a better understanding of our proposed method.

Normalizing flow. It is a widely-used method for unsupervised density estimation (Gudovskiy, Ishizaka, and Kozuka 2022; Liu, Tan, and Zhou 2022), which employs an invertible affine transformation to convert an original distribution into a target distribution. When direct density estimation based on the original data distribution \mathcal{X} is not feasible, alternatives to direct density estimation are to estimate the density of the target distribution \mathcal{Z} . We consider the case of a source sample $x \in \mathcal{R}^D$ drawn from the distribution \mathcal{X} and a target distribution sample $z \in \mathcal{R}^D$ drawn from the distribution \mathcal{Z} . $z = f_\theta(x)$ is a one-to-one mapping between \mathcal{X} and \mathcal{Z} . Utilizing the change of variable formula, we can

derive that

$$P_{\mathcal{X}}(x) = P_{\mathcal{Z}}(z) \left| \det \frac{\partial z}{\partial x^T} \right|. \quad (1)$$

Taking advantage of the invertibility of mapping functions and the tractability of Jacobian determinants $\left| \det \frac{\partial f_\theta(x)}{\partial x^T} \right|$. In flow models, the aim is to achieve $\hat{z} = z$, where $\hat{z} = f_\theta(x)$. When additional conditions C are incorporated, density estimation can be greatly improved. The conditional normalizing flow models represent the mapping as $z = f_\theta(x|C)$. The parameters θ of the $f_\theta(x)$ are optimized by maximum likelihood estimation (MLE):

$$\theta^* = \arg \max_{\theta} (\log(P_{\mathcal{Z}}(f_\theta(x|C))) + \log\left(\left| \det \frac{\partial f_\theta(x)}{\partial x^T} \right| \right)). \quad (2)$$

Notation and Problem Definition. Multivariate Time Series (MTS) are defined as $x = (x_1, x_2, \dots, x_k)$ where $x_i \in \mathcal{R}^L$. For each entity, k signifies the number of entities, while L signifies the number of observations of each entity. We use the z-score to normalize the time series of different entities

$$\bar{x}_i = \frac{x_i - \text{mean}(x_i)}{\text{std}(x_i)}, \quad (3)$$

where $\text{mean}(x_i)$ and $\text{std}(x_i)$ represent the mean and standard deviation of the entity i along the time dimension, respectively. A sliding window with size T and stride size S is used to sample the normalized MTS to preserve the temporal correlations of the original series. The training sample x^c can be obtained by adjusting T and S , where c is the number of samples. For simplicity, x^c stands for $x^{cS:cS+T}$.

Graph Mixture of Experts

Our model, depicted in Fig. 2, comprises independent expert networks dedicated to the intra-level aggregation of global-guided entity representations.

Feature Extraction. Following previous work (Dai and Chen 2022; Zhou et al. 2023) for a fair comparison, we

model the temporal variations of each entity by using the RNN model, *i.e.*, LSTM (Hochreiter and Schmidhuber 1997) and GRU (Cho et al. 2014).

$$H_k^t = \text{RNN}(x_k^t, H_k^{t-1}), \quad (4)$$

where H_k^t is the hidden RNN features for a window sequence of entity k , x_c at time $t \in [cS : cS + T)$. The combination of entities H_k^t for $\forall k$ forms the initial input H of our proposed Graph-MoE network.

Graph Construction. Due to the mutual and evolving dependence among entities, we utilize self-attention to learn a dynamic graph structure by treating entities at a certain time window of multivariate time series as graph nodes. For each input window sequence x^c , two linear transformations are performed, and then, the pairwise relationship e_{ij}^c at time window c between nodes i and j is determined as:

$$e_{ij}^c = (\phi_e^1(x_i^c)) \cdot (\phi_e^2(x_j^c))^\top, \quad (5)$$

where ϕ_e^1 and ϕ_e^2 are two linear transformations. To quantify the relationship between node i and j , the attention score a_{ij}^c of the matrix A^c is calculated by

$$a_{ij}^c = \frac{\exp(e_{ij}^c)}{\sum_{j=1}^K \exp(e_{ij}^c)}. \quad (6)$$

The attention matrix A^c inherently represents the mutual dependencies among the entities, which is treated as the adjacency matrix of the constructed graph, consequently. With time series inputs continually changing, A^c also evolves to capture the dynamic interdependence between them.

Graph Updating. To derive the temporal embedding H_t^l for the l -th GNN layer of all entities at t with incorporating the temporal features H_t^{l-1} of the last graph layer, a graph convolution operation is performed through the learned graph A^c . Inspired by previous work GANF (Dai and Chen 2022) and MTGFlow (Zhou et al. 2023), we also find that the history information of the node itself helps enhance temporal relationships of time series. Hence, the l -th GNN layer of spatio-temporal condition at time t is defined as:

$$H_t^l = \text{ReLU}(A^c H_t^{l-1} W_1 + H_t^{l-1} W_2) \cdot W_3, \quad (7)$$

where W_1 and W_2 are graph convolution and history information weights, respectively. W_3 is used to improve the expression ability of condition representations. Superscript l indicates the output of the l -th GNN layer. The spatio-temporal condition H_c^l for window c of the l -th GNN layer is the concatenation of H_t^l along the time axis.

Mixture of Experts. A mixture of experts (MoE) is proposed and introduced to explore whether distinct global temporal features may favor graph information encoded at different levels. A mixture of experts dynamically combines the outputs of sub-models known as ‘‘experts’’ via a weighting function known as the ‘‘router’’. The proposed mixture of experts is formulated as follows:

$$C = \sum_{l=1}^L R_l \cdot f_l(H^l) = \sum_{l=1}^L R_l \cdot \bar{H}^l, \quad (8)$$

where $f_l(\cdot)$ denotes the l -th expert function, \bar{H}^l is the expert-aligned entity features of the l -th graph layer, and $R = [R_1; R_2; \dots; R_L]$ is the routing weights for the L experts. For each entity x_i , since each GNN layer gathers information from directly connected entities, stacking multiple GNN layers enables the learning of neighborhood relationships between entities over long distances. Thus, each layer of our Graph-MoE has a specific expert network that independently captures features at different levels that focus on the different range of neighborhoods.

Our MoE takes the current temporal embedding H obtained from the RNN encoder as queries and the spatio-temporal condition H^l extracted from the l -th GNN layer as key-value pairs to aggregate node features within each layer. Formally, the cross-attention block operates as follows:

$$\tilde{H}^l = \text{LN}(\text{Attention}(H, H^l))|_{Q:H, \{K,V\}:H^l}, \quad (9)$$

where LN is the layer normalization operation. To enhance the capabilities of intra-layer feature representations, we employ the FFN and the ReLU activation function in the expert network. The expert-aligned entity features \bar{H}^l are:

$$\bar{H}^l = \text{FFN}_l(\tilde{H}^l) = \phi_l^2(\text{ReLU}(\phi_l^1(\tilde{H}^l))), \quad (10)$$

where ϕ_l^1 and ϕ_l^2 refer to the multi-layer perceptron (MLP). Then we respectively concatenate the outputs obtained from the attention mechanism, forming hierarchical global-guided temporal features $\tilde{\mathbf{H}}$:

$$\tilde{\mathbf{H}} = [\tilde{H}^1, \tilde{H}^2, \dots, \tilde{H}^L]. \quad (11)$$

For current time step t , $\tilde{\mathbf{H}}$ is specified as $\tilde{\mathbf{H}}_t$, which is then used for memory updates and dynamic routers generation.

Memory-augmented Routers

The memory-augmented router is designed to extract features from the current temporal slice while also capturing the characteristics of the entire historical data. These extracted features are then utilized by the module to predict the importance coefficients of experts of different time series.

To exploit the characteristics of memory, inspired by (Liu et al. 2022; Chen et al. 2020), we propose memory-augmented routers to enhance the weighted combination of GNNs’ multi-layer outputs. In doing so, the memory module uses a matrix to store the global historical temporal features. In the process of updating, the matrix is revised step-by-step incorporating the output from previous steps. Then, at time step t , the matrix from the previous step, M_{t-1} , is functionalized as the query, and its concatenations with the previous output serve as the key and value to feed the multi-head attention module. The entire process is formulated as follows:

$$\begin{aligned} Y &= [M_{t-1}; \tilde{\mathbf{H}}_t], \\ Z &= \text{Attention}(M_{t-1}, Y)|_{Q:M_{t-1}, \{K,V\}:Y}, \end{aligned} \quad (12)$$

where $[M_{t-1}; \tilde{\mathbf{H}}_t]$ is the row-wise concatenation of M_{t-1} and $\tilde{\mathbf{H}}_t$. Considering that the memory module is performed in a recurrent manner along with the estimation process,

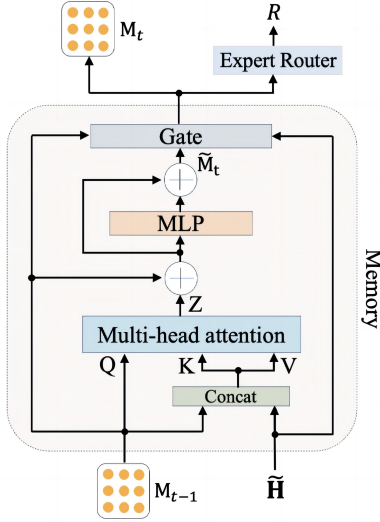


Figure 3: The framework of our memory-augmented routers.

it potentially suffers from gradient vanishing and exploding. We therefore introduce residual connections and a gate mechanism. The former is formulated as

$$\tilde{M}_t = \phi_M(Z + M_{t-1}) + Z + M_{t-1}, \quad (13)$$

where ϕ_M refers to the multi-layer perceptron (MLP). The detailed structure of the gate mechanism in the memory module is shown in Fig. 3, where the forget and input gates are applied to balance the inputs from M_{t-1} and \tilde{H}_t , respectively. The forget and input gates are formalized as

$$G_t^f = \tilde{H}_t W^f + \tanh(M_{t-1}) \cdot U^f, \quad (14)$$

$$G_t^i = \tilde{H}_t W^i + \tanh(M_{t-1}) \cdot U^i, \quad (15)$$

where W^f and W^i are trainable weights for \tilde{H}_t in each gate; and similarly, U^f and U^i are the trainable weights for M_{t-1} in each gate. The final output of the gate mechanism is formalized as follows:

$$M_t = \sigma(G_t^f) \odot M_{t-1} + \sigma(G_t^i) \odot \tanh(\tilde{M}_t), \quad (16)$$

where \odot refers to the Hadamard product, σ is the sigmoid activation, and M_t is the output of the memory module at step t , which contains both the global historical temporal information and the current temporal features. We then update the memory by replacing M_{t-1} with M_t in the next step.

The output $\phi_R(M_t)$, derived from the memory module, is then fed into the expert router, which is a linear transformation ϕ_R of the input followed by a softmax layer:

$$R = \text{Softmax}(\phi_R(M_t)), \quad (17)$$

where the output R preserves the predicted expert weights through the memory-augmented router. Finally, we leverage Eq. (8) to generate the spatio-temporal condition C .

Experiment

Experiment Setup

Datasets. Following GANF (Dai and Chen 2022) and MTGFlow (Zhou et al. 2023), we conduct experiments on five

widely-used public datasets for MTS anomaly detection. The details are illustrated as follows: SWaT (Mathur and Tippenhauer 2016) consists of 51 sensors from an industrial water treatment plant, including 41 attacks over four days. WADI (Ahmed, Palleti, and Mathur 2017) is a collection of 123 sensor and actuator data from the WADI testbed, at the frequency of one second. PSM (Abdulaal, Liu, and Lancewicki 2021) is a collection of multiple application server nodes at eBay with 25 features, providing ground truths created by experts over 8 weeks. MSL (Hundman et al. 2018) is a collection of the sensor and actuator data of the Mars rover with 55 dimensions. SMD (Su et al. 2019) is a 5-week-long dataset that is collected from a large Internet company with 38 dimensions.

Evaluation Metrics. Following previous studies (Dai and Chen 2022; Zhou et al. 2023), Graph-MoE is designed to detect anomalies at the window level, where a window is labeled as abnormal if any time point within it is identified as anomalous. The performance of the model is evaluated using the Area Under the Receiver Operating Characteristic Curve (AUROC).

Implementation Details. We set the window size as 60 and the stride size as 10. During the training stage, we utilize an Adam optimizer. For the SWaT dataset, we use a single flow block with a batch size of 640 and a learning rate of 0.005. The batch size of the WADI dataset is 320, and the learning rate is 0.0035. Other datasets employ two flow blocks with a batch size of 256 and a learning rate of 0.002. We follow the dataset setting of previous work (Zhou et al. 2023; Dai and Chen 2022) and split the original testing dataset by 60% for training, 20% for validation, and 20% for testing in SWaT. For other datasets, the training split contains 60% data, and the test split contains 40% data. All our experiments are conducted on GPU, running for 80 epochs.

Baselines. To demonstrate the superiority of our method, we compare it with current competitive approaches, including the state-of-the-art methods that encompass both semi-supervised and unsupervised approaches to anomaly detection. Notably, semi-supervised methods include DeepSAD (Ruff et al. 2020) and DROCC (Goyal et al. 2020). Besides, unsupervised methods include DeepSVDD (Ruff et al. 2018), ALOCC (Sabokrou et al. 2020), USAD (Audibert et al. 2020), DAGMM (Zong et al. 2018), GANF (Dai and Chen 2022) and MTGFlow (Zhou et al. 2023).

Main Comparison

As shown in Table 1, we compare Graph-MoE with the aforementioned baselines and report the average performance and standard deviations on AUROC scores through five random repetitions. The results show that our Graph-MoE reaches an exceptionally high AUROC score compared to the other eight baselines. Note that the standard deviation observed in the SMD arises from the fact that it encompasses 28 distinct sub-datasets. For each sub-dataset, we evaluate the performance and subsequently calculate the average of all these individual results. From the results of Table 1, we have the following observations.

Method	Venue	Datasets				
		SWaT	WADI	PSM	MSL	SMD
DeepSVDD (Ruff et al. 2018)	ICML2018	66.8±2.0	83.5±1.6	67.5±1.4	60.8±0.4	75.5±15.5
DAGMM (Zong et al. 2018)	ICLR 2018	72.8±3.0	77.2±0.9	64.6±2.6	56.5±2.6	78.0±9.2
ALOCC (Sabokrou et al. 2020)	TNNLS 2020	77.1±2.3	83.3±1.8	71.8±1.3	60.3±0.9	80.5±11.1
DROCC (Goyal et al. 2020)	ICML 2020	72.6±3.8	75.6±1.6	74.3±2.0	53.4±1.6	76.7±8.7
DeepSAD (Ruff et al. 2020)	ICLR 2020	75.4±2.4	85.4±2.7	73.2±3.3	61.6±0.6	85.9±11.1
USAD (Audibert et al. 2020)	KDD 2020	78.8±1.0	86.1±0.9	78.0±0.2	57.0±0.1	86.9±11.7
GANF (Dai and Chen 2022)	ICLR 2022	79.8±0.7	90.3±1.0	81.8±1.5	64.5±1.9	89.2±7.8
MTGFlow (Zhou et al. 2023)	AAAI 2023	84.8±1.5	91.9±1.1	85.7±1.5	67.2±1.7	91.3±7.6
Ours (Graph-MoE)		87.2±1.3	94.2±0.8	88.0±0.7	72.1±1.1	93.3±5.6

Table 1: Comparison with the state-of-the-art methods in anomaly detection on five challenging datasets, *i.e.*, SWaT, WADI, PSM, MSL, and SMD. The best results are highlighted in bold.

# of experts	Datasets	
	SWaT	WADI
1	86.2±0.7	92.5±1.2
2	86.8±0.4	93.4±0.6
3	87.2±1.3	94.2±0.8
4	85.6±2.1	92.6±0.7

Table 2: The results of ablation studies on the SWaT dataset to discuss the number of experts, which is the important hyperparameter of our method.

- Compared with semi-supervised methods, our method achieves significant improvements on metric AUROC%. For instance, on the SWaT dataset, our method achieves improvements over DeepSAD and DROCC, +15.6% and +20.1%, respectively. The reason for such remarkable improvements is that, unlike our Graph-MoE, DROCC projects all training samples into a hypersphere, hindering their ability to accurately distinguish between normal and abnormal samples. DeepSAD relies on the distance to the center to detect anomalies, and its effectiveness is limited when handling complex anomaly patterns.
- Our model also achieves better performance than the unsupervised methods in the baseline. For example, on the WADI dataset, our model demonstrate a +22.0% increase compared to DAGMM, +4.3% compared to GANF, and +2.5% compared to MTGFlow. The reason is that our Graph-MoE leverages the multi-layer GNN to comprehensively capture information from short-distance to long-distance neighborhoods, enhancing the understanding of complex relationships in multivariate time series. By introducing the Mixture of Experts (MoE) and the Memory-augmented Routers (MAR), the model adaptively selects suitable experts based on spatio-temporal features, effectively combining global and local information for precise detection.

Ablation Studies

Effectiveness on the number of Graph-MoE layers. In our proposed method Graph-MoE, the most critical parameter is the number of layers. We explore the influence of different layers of Graph-MoE on anomaly detection performance,

Components		SWaT	WADI
MoE	MAR		
×	×	85.5±1.1	92.2±0.2
×	✓	86.2±1.2	92.9±1.0
✓	×	86.8±1.2	93.5±0.4
✓	✓	87.2±1.3	94.2±0.8

Table 3: The results of ablation studies on SWaT and WADI datasets to discuss the effectiveness of our proposed components. MoE and MAR stand for the mixture of experts and memory-augmented routers, respectively.

and the results are shown in Table 2. We can observe that the AUROC fraction increases gradually with the increase in the number of layers and reaches a peak at the third-layer Graph-MoE. However, as the number of layers increases to 4, the performance declines.

The reason is that when the number of layers is set to 1 and 2, the model can only aggregate a restricted amount of neighborhood information, which is insufficient for capturing complex long-distance relationships. Moreover, an excessive number of layers can lead to the occurrence of over-smoothing, which reduces the distinctiveness between node features, diminishes diversity and ultimately weakens the model’s representation ability. The results show that a reasonable selection of layers can significantly improve the detection ability of the model.

Discussion on each proposed module. To further investigate the contribution of each component in our method, we conduct a set of ablation studies. The results are shown in Table 3. In this table, “×” indicates that the module is not used, while “✓” indicates that the module is utilized. The components MoE and MAR represent the mixture of experts and memory-augmented routers, respectively. Initially, without the inclusion of MAR and MoE, there are modest performance improvements of 0.8% and 0.3% on SWaT and WADI over MTGFlow, respectively, which suggests that the base graph structure alone is somewhat effective in capturing spatio-temporal information. However, when MAR and MoE are introduced separately, both contribute to further performance gains, with MoE having a more pronounced

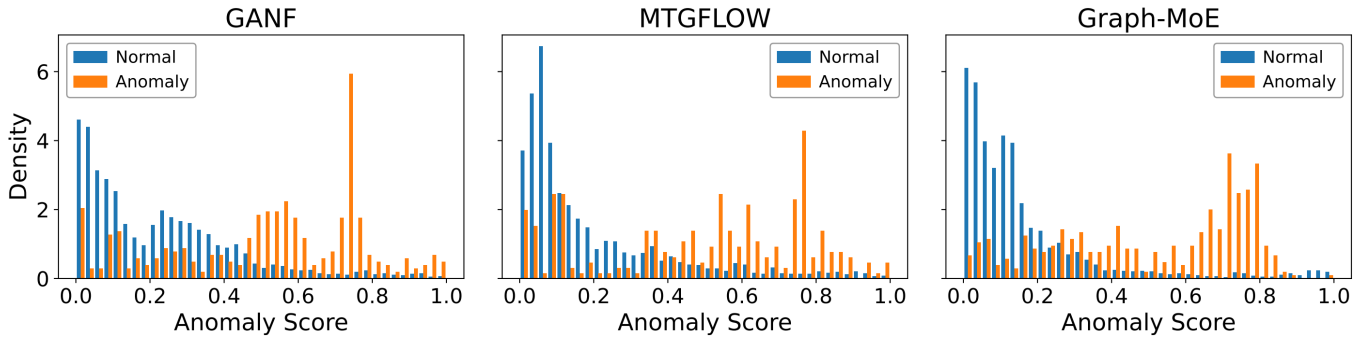


Figure 4: Distribution of anomaly scores for normal and anomaly data across GANF, MTGFlow, and Graph-MoE models.

Methods	SWaT	Improvement($\Delta\%$)
GANF	79.8 \pm 0.7	
GANF+Graph-MoE	82.6\pm0.7	$\Delta\%=3.5$
MTGFlow	84.8 \pm 1.5	
MTGFlow+Graph-MoE	87.2\pm1.3	$\Delta\%=2.8$
USD	90.2 \pm 0.9	
USD+Graph-MoE	92.3\pm1.4	$\Delta\%=2.3$

Table 4: We integrate our Graph-MoE into the three baseline methods of GANF, MTGFlow, and USD, and the results show the superiority of our model.

impact. The results could be attributed to MoE’s ability to better integrate expert knowledge from different GNN layers, enhancing the model’s representational capacity. When MAR and MoE are both introduced into our model, their synergistic effect leads to substantial performance improvements of 2.8% on the SWaT dataset and 2.5% on the WADI dataset, which underscores the effectiveness of combining MAR’s global historical feature extraction with MoE’s integration of multi-level information in addressing complex time series anomaly detection challenges.

Discussion on the plug-and-play properties. To evaluate the plug-and-play capability of our model, we select three representative graph-based anomaly detection methods, namely GANF (Dai and Chen 2022), MTGFlow (Zhou et al. 2023), and USD (Liu et al. 2024a). As shown in Table 4, we apply Graph-MoE upon these three baselines. Specifically, after incorporating Graph-MoE, GANF’s performance increases from 79.8 to 82.6, representing a gain of 3.5 percentage points, the largest improvement observed. MTGFlow’s performance rises from 84.8 to 87.2, with a 2.8% increase. Similarly, USD’s performance improves from 90.2 to 92.3, an increase of 2.3 percentage points. This consistent performance enhancement proves that Graph-MoE effectively strengthens anomaly detection capabilities across different graph-based models, demonstrating the strong plug-and-play capabilities of our proposed Graph-MoE.

Qualitative Results

Moreover, we visualize the distributions of data on the SWaT dataset to examine the effectiveness of our method. To investigate the anomaly discrimination ability of Graph-

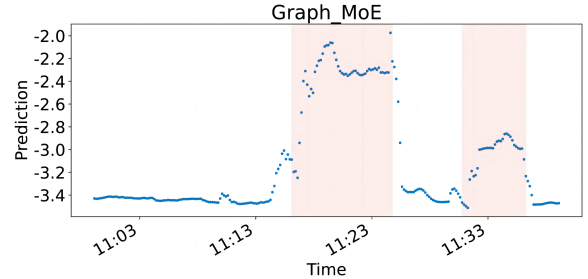


Figure 5: Anomaly detection over time by our Graph-MoE.

MoE, we present the normalized anomaly score for GANF, MTGFlow and Graph-MoE in Fig. 4. As it is displayed, for normal series, anomaly scores of Graph-MoE are more centered at 0 than those of GANF and MTGFlow, and the overlap areas of normal and abnormal scores are also smaller in Graph-MoE, reducing the false positive ratio. This larger score discrepancy corroborates that Graph-MoE has superior detection performance. Fig. 5 shows the point plot of Graph-MoE anomalies predicted outputs over time. The X-axis displays the progression of time, while the Y-axis indicates the log-likelihood of the anomaly. Abnormal series are highlighted with a red background. This fluctuation in log likelihoods confirms that our proposed Graph-MoE has capabilities for detecting anomalies by identifying low-density regions within the modeled distribution.

Conclusion

In this work, we proposed Graph-MoE, an unsupervised-learning based method for anomaly detection in multivariate time series data. Unlike previous methods, we employ a multi-layer, multi-expert learning architecture to dynamically model the complex interdependencies between entities and the various inherent characteristics of each entity at different levels, capturing both local and global information and enhancing feature representations of each entity. We also propose Memory-augmented Routers, which leverage the memory mechanism to store the patterns of global historical temporal features to predict the importance coefficients of experts, therefore facilitating effective inter-level aggregation. Experimental results demonstrate that the proposed framework achieves state-of-the-art performance.

Acknowledgements

This work is supported by the National Key Research and Development Program under Grant No. 2023YFC3303800, and the National Science Fund for Excellent Young Scholars under Grant 62222212.

References

- Abdulaal, A.; Liu, Z.; and Lancewicki, T. 2021. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *KDD*, 2485–2494.
- Ahmed, C. M.; Palleti, V. R.; and Mathur, A. P. 2017. 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*, 25–28.
- Audibert, J.; Michiardi, P.; Guyard, F.; Marti, S.; and Zuluaga, M. A. 2020. Usad: Unsupervised anomaly detection on multivariate time series. In *KDD*, 3395–3404.
- Cai, L.; Chen, Z.; Luo, C.; Gui, J.; Ni, J.; Li, D.; and Chen, H. 2021. Structural temporal graph neural networks for anomaly detection in dynamic graphs. In *ACM CIKM*, 3747–3756.
- Chen, S.; and Wang, J. 2022. Heterogeneous interaction modeling with reduced accumulated error for multiagent trajectory prediction. *IEEE Transactions on Neural Networks and Learning Systems*.
- Chen, S.; Wang, J.; and Li, G. 2021. Neural relational inference with efficient message passing mechanisms. In *AAAI*, volume 35, 7055–7063.
- Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating Radiology Reports via Memory-driven Transformer. In *EMNLP*, 1439–1449.
- Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha, Qatar: Association for Computational Linguistics.
- Choi, K.; Yi, J.; Park, C.; and Yoon, S. 2021. Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE access*, 9: 120043–120065.
- Dai, E.; and Chen, J. 2022. Graph-Augmented Normalizing Flows for Anomaly Detection of Multiple Time Series. In *ICLR*.
- Deng, A.; and Hooi, B. 2021. Graph neural network-based anomaly detection in multivariate time series. In *AAAI*, volume 35, 4027–4035.
- Gao, J.; Chen, M.; and Xu, C. 2023. Vectorized Evidential Learning for Weakly-supervised Temporal Action Localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12): 15949 – 15963.
- Gao, J.; Zhang, T.; and Xu, C. 2020. Learning to model relationships for zero-shot video classification. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3476–3491.
- Goyal, S.; Raghunathan, A.; Jain, M.; Simhadri, H. V.; and Jain, P. 2020. DROCC: Deep robust one-class classification. In *ICML*, 3711–3721. PMLR.
- Gudovskiy, D.; Ishizaka, S.; and Kozuka, K. 2022. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 98–107.
- Guo, D.; Li, K.; Hu, B.; Zhang, Y.; and Wang, M. 2024. Benchmarking Micro-action Recognition: Dataset, Method, and Application. *IEEE Transactions on Circuits and Systems for Video Technology*.
- He, J.; Xu, Q.; Jiang, Y.; Wang, Z.; and Huang, Q. 2024. ADA-GAD: Anomaly-Denoised Autoencoders for Graph Anomaly Detection. In *AAAI*, volume 38, 8481–8489.
- Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.
- Hu, Y.; Gao, J.; Dong, J.; Fan, B.; and Liu, H. 2023. Exploring rich semantics for open-set action recognition. *IEEE Transactions on Multimedia*.
- Hundman, K.; Constantinou, V.; Laporte, C.; Colwell, I.; and Soderstrom, T. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *KDD*, 387–395.
- Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; and Chen, Z. 2021. {GS}hard: Scaling Giant Models with Conditional Computation and Automatic Sharding. In *ICLR*.
- Lewis, M.; Bhosale, S.; Dettmers, T.; Goyal, N.; and Zettlemoyer, L. 2021. Base layers: Simplifying training of large, sparse models. In *ICML*, 6265–6274. PMLR.
- Liu, B.; Tan, P.-N.; and Zhou, J. 2022. Unsupervised anomaly detection by robust density estimation. In *AAAI*, volume 36, 4101–4108.
- Liu, H.; Qiu, X.; Shi, Y.; and Zang, Z. 2024a. USD: Unsupervised Soft Contrastive Learning for Fault Detection in Multivariate Time Series. *arXiv preprint arXiv:2405.16258*.
- Liu, X.; Hu, C.; Zhang, R.; Sun, K.; Mensah, S.; and Mao, Y. 2024b. Multimodal Relation Extraction via a Mixture of Hierarchical Visual Context Learners. In *Proceedings of the ACM on Web Conference 2024*, 4283–4294.
- Liu, X.; Lyu, X.; Zhang, X.; Gao, J.; and Chen, J. 2022. Memory augmented graph learning networks for multivariate time series forecasting. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 4254–4258.
- Ma, C.; Ma, L.; Zhang, Y.; Sun, J.; Liu, X.; and Coates, M. 2020. Memory augmented graph neural networks for sequential recommendation. In *AAAI*, volume 34, 5045–5052.
- Ma, J.; Zhao, Z.; Yi, X.; Chen, J.; Hong, L.; and Chi, E. H. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *KDD*, 1930–1939.

- Ma, K.; Xu, Q.; Zeng, J.; Cao, X.; and Huang, Q. 2021. Poisoning attack against estimating from pairwise comparisons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6393–6408.
- Ma, K.; Xu, Q.; Zeng, J.; Li, G.; Cao, X.; and Huang, Q. 2022. A tale of hodgerank and spectral method: Target attack against rank aggregation is the fixed point of adversarial game. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4090–4108.
- Ma, K.; Xu, Q.; Zeng, J.; Liu, W.; Cao, X.; Sun, Y.; and Huang, Q. 2024. Sequential manipulation against rank aggregation: theory and algorithm. *IEEE transactions on pattern analysis and machine intelligence*.
- Mathur, A. P.; and Tippenhauer, N. O. 2016. 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)*, 31–36. IEEE.
- Rasul, K.; Seward, C.; Schuster, I.; and Vollgraf, R. 2021a. Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting. In *ICML*, 8857–8868. PMLR.
- Rasul, K.; Sheikh, A.-S.; Schuster, I.; Bergmann, U. M.; and Vollgraf, R. 2021b. Multivariate Probabilistic Time Series Forecasting via Conditioned Normalizing Flows. In *ICLR*.
- Ren, H.; Xu, B.; Wang, Y.; Yi, C.; Huang, C.; Kou, X.; Xing, T.; Yang, M.; Tong, J.; and Zhang, Q. 2019. Time-series anomaly detection service at microsoft. In *KDD*, 3009–3017.
- Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S. A.; Binder, A.; Müller, E.; and Kloft, M. 2018. Deep one-class classification. In *ICML*, 4393–4402. PMLR.
- Ruff, L.; Vandermeulen, R. A.; Görnitz, N.; Binder, A.; Müller, E.; Müller, K.-R.; and Kloft, M. 2020. Deep Semi-Supervised Anomaly Detection. In *ICLR*.
- Sabokrou, M.; Fathy, M.; Zhao, G.; and Adeli, E. 2020. Deep end-to-end one-class classifier. *IEEE transactions on neural networks and learning systems*, 32(2): 675–684.
- Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *ICLR*.
- Su, Y.; Zhao, Y.; Niu, C.; Liu, R.; Sun, W.; and Pei, D. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *KDD*, 2828–2837.
- Tan, Q.; Zhang, J.; Liu, N.; Huang, X.; Yang, H.; Zhou, J.; and Hu, X. 2021. Dynamic memory based attention network for sequential recommendation. In *AAAI*, volume 35, 4384–4392.
- Tian, Y.; Chen, G.; and Song, Y. 2021. Aspect-based sentiment analysis with type-aware graph convolutional networks and layer ensemble. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2910–2922.
- Wang, H.; Jiang, Z.; You, Y.; Han, Y.; Liu, G.; Srinivasa, J.; Kompella, R.; Wang, Z.; et al. 2024a. Graph mixture of experts: Learning on large-scale graphs with explicit diversity modeling. *Advances in Neural Information Processing Systems*, 36.
- Wang, W.; Dong, L.; Cheng, H.; Liu, X.; Yan, X.; Gao, J.; and Wei, F. 2024b. Augmenting language models with long-term memory. *Advances in Neural Information Processing Systems*, 36.
- Wu, Z.; Pan, S.; Long, G.; Jiang, J.; Chang, X.; and Zhang, C. 2020. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *KDD*, 753–763.
- Wu, Z.; Pan, S.; Long, G.; Jiang, J.; and Zhang, C. 2019. Graph WaveNet for Deep Spatial-Temporal Graph Modeling. In *IJCAI*, 1907–1913. International Joint Conferences on Artificial Intelligence Organization.
- Xu, K.; Li, C.; Tian, Y.; Sonobe, T.; Kawarabayashi, K.-i.; and Jegelka, S. 2018. Representation learning on graphs with jumping knowledge networks. In *ICML*, 5453–5462. PMLR.
- Xu, N.; Gao, Y.; Liu, A.-A.; Tian, H.; and Zhang, Y. 2024a. Multi-Modal Validation and Domain Interaction Learning for Knowledge-based Visual Question Answering. *TKDE*.
- Xu, N.; Gao, Y.; Zhang, T.-T.; Tian, H.; and Liu, A.-A. 2024b. Cross-Modal Coherence-Enhanced Feedback Prompting for News Captioning. In *ACM MM*, 9369–9377.
- Xu, N.; Zhang, T.; Tian, H.; and Liu, A.-A. 2024c. Rule-driven News Captioning. *IEEE Transactions on Circuits and Systems for Video Technology*.
- Yang, L.; Hu, W.; Xu, J.; Shi, R.; He, D.; Wang, C.; Cao, X.; Wang, Z.; Niu, B.; and Guo, Y. 2024a. GAUSS: GrAph-customized Universal Self-Supervised Learning. In *Proceedings of the ACM on Web Conference 2024*, 582–593.
- Yang, L.; Shi, R.; Zhang, Q.; Wang, Z.; Cao, X.; Wang, C.; et al. 2024b. Self-supervised graph neural networks via low-rank decomposition. *Advances in Neural Information Processing Systems*, 36.
- Zhang, W.; Zhang, C.; and Tsung, F. 2022. GRELEN: Multivariate Time Series Anomaly Detection from the Perspective of Graph Relational Learning. In *IJCAI*, 2390–2397.
- Zhao, H.; Wang, Y.; Duan, J.; Huang, C.; Cao, D.; Tong, Y.; Xu, B.; Bai, J.; Tong, J.; and Zhang, Q. 2020. Multivariate time-series anomaly detection via graph attention network. In *ICDM*, 841–850. IEEE.
- Zhou, Q.; Chen, J.; Liu, H.; He, S.; and Meng, W. 2023. Detecting multivariate time series anomalies with zero known label. In *AAAI*, volume 37, 4963–4971.
- Zong, B.; Song, Q.; Min, M. R.; Cheng, W.; Lumezanu, C.; Cho, D.; and Chen, H. 2018. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *ICLR*.