Information Fusion 89 (2023) 527–536

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

MST-GAT: A multimodal spatial–temporal graph attention network for time series anomaly detection

Chaoyue Ding [a], Shiliang Sun [a,b], Jing Zhao [a,*]

[a] School of Computer Science and Technology, East China Normal University, 3663 North Zhongshan Road, Shanghai 200062, PR China
[b] College of Mathematics and Computer Science, Zhejiang Normal University, 688 Yingbin Road, Jinhua 321004, PR China

ARTICLE INFO

Keywords:
Multimodal time series
Anomaly detection
Graph attention networks
Unsupervised learning

ABSTRACT

Multimodal time series (MTS) anomaly detection is crucial for maintaining the safety and stability of working devices (e.g., water treatment system and spacecraft), whose data are characterized by multivariate time series with diverse modalities. Although recent deep learning methods show great potential in anomaly detection, they do not explicitly capture spatial–temporal relationships between univariate time series of different modalities, resulting in more false negatives and false positives. In this paper, we propose a multimodal spatial–temporal graph attention network (MST-GAT) to tackle this problem. MST-GAT first employs a multimodal graph attention network (M-GAT) and a temporal convolution network to capture the spatial–temporal correlation in multimodal time series. Specifically, M-GAT uses a multi-head attention module and two relational attention modules (i.e., intra- and inter-modal attention) to model modal correlations explicitly. Furthermore, MST-GAT optimizes the reconstruction and prediction modules simultaneously. Experimental results on four multimodal benchmarks demonstrate that MST-GAT outperforms the state-of-the-art baselines. Further analysis indicates that MST-GAT strengthens the interpretability of detected anomalies by locating the most anomalous univariate time series.

1. Introduction

Anomaly detection has gained much attention in different fields (e.g., images [1], text [2], time series [3], etc.), aiming to find instances that deviate significantly from all other observations [4,5]. In this paper, we focus on multimodal time series (MTS) anomaly detection, which is the subtask of anomaly detection. MTS anomaly detection is commonly used to monitor diverse modalities (e.g., temperature, speed, and power) of sensors in industrial devices and information technology systems (i.e., entities), and the data stream from each sensor is seen as a univariate time series. Multimodal time series data can facilitate the detection of complex anomalies, which is not obvious when monitoring each modality independently [6]. Furthermore, before the entity is partially or fully disrupted, the timely detection of anomalies helps the user to troubleshoot.

Conventionally, experienced engineers manually create static thresholds for each monitored time series to perform anomaly detection. However, with the exponential growth of data size in recent years, this manual approach will be labor-intensive [7]. Moreover, determining an optimal threshold for each sensor is challenging, especially when multimodal sensing is employed in the entity. Many anomaly detection methods have been proposed to solve the above problem, and they combine the anomaly detection results of all univariate time series in an entity to detect anomalies [3,8]. However, an entity of multimodal time series often involves massively interconnected univariate time series, which continuously generate multimodal time series data, and these sensor data are typically correlated in complex non-linear ways. Thus, a single univariate time series fails to respond to the overall state of the entity, and methods that simply combine the detection results of multiple univariate time series tend to perform poorly. MTS anomaly detection has been challenging due to the complex spatial dependence (e.g., topological structure and modal correlation) and temporal dependence (e.g., period and trend) of multimodal time series. Besides, multimodal time series contains not only correlations between time series from the same modality (referred to as intra-modal correlations) but also correlations between time series from different modalities (referred to as inter-modal correlations).

Previous methods for MTS anomaly detection take temporal dependence into account, including support vector regression [9], Bayesian models [10], autoregressive integrated moving average (ARIMA) [11] and recurrent neural network (RNN)-based models [12]. These methods can capture the dynamic changes in temporal dimension but ignore the spatial dependence between different time series. To remedy this, some

* Corresponding author.
E-mail address: jzhao@cs.ecnu.edu.cn (J. Zhao).

https://doi.org/10.1016/j.inffus.2022.08.011
Received 20 December 2021; Received in revised form 6 July 2022; Accepted 9 August 2022
Available online 13 August 2022

researchers introduced convolutional neural networks (CNNs) to better model spatial relationships [3]. However, CNNs are usually applied to regular data such as image video and speech data, which would lead to inferior performance in graph data due to the complex topology among multimodal time series. Graph neural networks (GNNs) [13] are more effective paradigms to build complex topological relationships in graph data, which have also been developed for anomaly detection and achieved promising results. Specifically, Hang et al. [14] adopted graph neural networks and gated recurrent unit (GRU) [15] to study the spatial–temporal relationships of time series. Although the previous methods have made fruitful progress, they fail to explicitly capture the multimodal correlation among multimodal time series.

Another challenge is to provide interpretation for anomaly detection results. To provide users with more valuable information, MST-GAT interprets each anomaly by locating a univariate time series that is most likely to cause this anomaly. The reconstruction probability of the reconstruction-based method is usually used to interpret detected anomalies. For instance, OmniAnomaly [16] leveraged variational autoencoder (VAE) [17] to produce reconstruction probabilities for each time series, which is used to explain the detected anomalies. Although these methods can capture the stochasticity of entire time series, researches show that they fail to perform well on periodic scenarios, while prediction-based methods can overcome this deficiency [14].

In this paper, we propose the multimodal spatial–temporal graph attention network, termed MST-GAT, which adopts the prevalent graph attention networks (GATs) [18] to explicitly capture modal dependencies between multimodal time series. More specifically, we design the multimodal graph attention network (M-GAT), which includes a multi-head attention module and two relational attention modules, i.e., intra- and inter-modal attention, to capture the spatial dependencies between multimodal time series. Explicitly modeling different relationships in multimodal time series is conducive to obtaining a better feature representation of input data. We then introduce a temporal convolution network to capture temporal dependencies in each time series with a standard convolution operation on time slices. Additionally, we jointly optimize a reconstruction module and a prediction module to integrate their advantages. The reconstruction module accounts for reconstructing the input data, while the prediction module aims at predicting the feature of the next timestamp. The reconstruction probability and the prediction error are further used to explain the detected anomalies.

The main contributions of this paper are summarized as follows:

- We propose MST-GAT, a novel MTS anomaly detection method based on graph attention networks. To the best of our knowledge, MST-GAT pioneers the exploration of explicitly modeling spatial–temporal dependencies in multimodal time series data for anomaly detection.
- We jointly optimize a variational autoencoder-based reconstruction module and a multilayer perceptron (MLP)-based prediction module to integrate their advantages. MST-GAT achieves the highest F1-score, all above 0.60, the best AUC, all above 0.92, outperforming the strong baselines on benchmark datasets. Ablation studies further prove the effectiveness of the different modules in the MST-GAT.
- We devise an efficient anomaly interpretation method for MTS anomaly detection based on the reconstruction and prediction results. MST-GAT is well interpretable and is capable of obtaining results that are consistent with human intuition.

The rest of the paper is structured as follows. Section 2 presents the related works. Section 3 introduces the proposed MST-GAT. In Section 4, experimental results demonstrate the effectiveness of the proposed method. Finally, conclusions and future research directions are given in the last section.

2. Related works

This section briefly introduces related works, including time series anomaly detection, graph neural networks, and multimodal machine learning.

2.1. Time series anomaly detection

Time series anomaly detection has been investigated for decades, and various types of approaches have been proposed [19,20]. Traditional anomaly detection approaches can be classified into clustering [21], distance-based [22], density-based [23], and isolation-based [24] methods. Recently, deep learning approaches have attracted much attention due to the powerful representational capabilities of deep neural networks [25]. Existing deep learning methods can be categorized into two paradigms, namely reconstruction-based and prediction-based methods.

A reconstruction-based method learns the potential distribution of the entire time series. Deep autoencoding Gaussian model (DAGMM) [26] obtained low-dimensional features and reconstruction-based anomaly scores by combining a deep autoencoder network and a Gaussian mixture model (GMM). OmniAnomaly [16] employed VAE into an end-to-end structure to reconstruct the input data, and it detected anomalies according to the reconstruction probability. RAMED [27] utilized a multi-resolution network to encourage reconstructed outputs to match the global temporal shape of input. A prediction-based method predicts the value of the following timestamp and anomalies according to the prediction residual. Hundman et al. [28] demonstrated the feasibility of long short-term memory (LSTM) in detecting spacecraft anomalies and introduced an approach for setting thresholds dynamically without relying on annotations. Graph deviation network (GDN) [29] utilized graph attention networks (GATs) to perform structure learning in multivariate time series and interpreted a detected anomaly by attention weights.

Previous works have proved that reconstruction-based and prediction-based methods are complementary in different scenarios [14]. Therefore, we propose a joint network to integrate the advantages of these two paradigms. Nevertheless, none of the existing methods consider explicitly capturing the relationship between multimodal data. Our work aims to address this problem by using multimodal graph attention to explicitly construct spatial–temporal dependencies within multimodal time series.

2.2. Graph neural networks

Graph neural networks (GNNs) have gained remarkable success in graph structure data such as social networks [30] and medical science [31]. Typical GNNs supposes that the representation of a node is affected by its neighboring nodes in a graph structure. Graph convolutional networks (GCNs) include spectral methods and spatial methods [32]. Spectral methods suffer from the drawback of basic dependence, and spatial methods are limited by the lack of shift-invariance [33,34].

Attention mechanisms have become an effective and widely used component of the sequence-to-sequence models in many deep learning applications [35,36]. Recently, attention mechanisms have been introduced to graph neural networks. Graph attention networks (GATs) [18] utilized the attention mechanisms to assign aggregation weights to neighboring nodes. Relevant variants of graph attention networks have made progress in tasks related to time series modeling, e.g., traffic flow forecasting [37] and time series forecasting [38]. Graph attention networks can better extract the spatial feature and shows superior performances than graph convolutional neural network in the directed graph [39]. GATs map the input feature vectors h into the aggregated

representation h' using the attention mechanisms. The attention score a_{ij} is formulated as:

$$a_{ij} = \frac{\exp\left(\hat{a}_{ij}\right)}{\sum_{k\in\mathcal{N}_i}\exp\left(\hat{a}_{ik}\right)}, \tag{1}$$

$$\hat{a}_{ij} = \text{LeakyReLU}(\pi\left(\mathbf{W}h_i, \mathbf{W}h_j\right)), \tag{2}$$

where $\mathcal{N}_i$ is the neighbor set of node i, $\hat{a}_{ij}$ denotes the attention score between node i and node j before normalization, $\pi(\cdot)$ represents the correlation function between nodes, $\mathbf{W}$ is the weight matrix, h_i is the feature representation of node i, and LeakyReLU is an activation function. The output feature of each node can be calculated as:

$$h'_i = \sigma\left(\sum_{j\in\mathcal{N}_i} a_{ij}\mathbf{W}h_j\right), \tag{3}$$

where σ denotes the sigmoid activation function.

2.3. Multimodal machine learning

Multimodal machine learning aims to take full advantage of information from different modalities [40,41]. Specifically, multimodal machine learning helps to integrate the complementary information across the multimodal data and facilitates extracting similar information to improve model robustness. Compared with models using single modal data, models using multimodal data always perform better. How to fuse multimodal information into a unified representation is a primary challenge. Multimodal machine learning can be realized by multi-kernel learning models [42], probabilistic graphical models [43], neural network models [44], etc. Among them, neural networks models have enabled significant performance improvements in many tasks due to their excellent ability to fuse multimodal data. For instance, Iwana et al. [45] used multimodal CNNs with the local distance-based representation to perform time series classification. Yang et al. [46] introduced an adaptive weighting algorithm and a multi-head co-attention network to model the association between textual and visual representations in multimodal machine translation.

Recently, multimodal machine learning has been introduced into anomaly detection [47]. Park et al. [6] trained a hidden Markov model (HMM) with multimodal data from non-anomalous scenarios and performed anomaly detection for robot manipulation. Park et al. [48] leveraged an LSTM-based autoencoder to detect anomalies of an assisted feeding robot.

3. Methodology

This section mainly introduces MST-GAT. MST-GAT uses the multimodal graph attention network and temporal convolution network to capture both temporal and spatial dependencies in multimodal time series.

3.1. Problem definition

Multimodal time series consist of time series from multiple modalities belonging to the same entity, and each modality can contain one or more time series. The MTS anomaly detection model is designed to detect anomalies at the timestamp level. Time series anomaly detection is usually regarded as an unsupervised task, and we assume that there are no anomalies during the training phase. We formulate this problem as follows. In the training phase, we train our model in the training set. The multimodal time series data are composed of N univariate time series with T timestamps, i.e., $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T] \in \mathbb{R}^{N\times T}$, and each univariate time series includes M modalities ($1 \le M \le N$). An observation the multimodal time series at time t ($t \le T$) is denoted by $\mathbf{x}_t = [x_{1,t}^{m_1}, x_{2,t}^{m_2}, \dots, x_{N,t}^{m_N}]^\top$, where $x_{i,t}^{m_i}$ ($m_i \in \{1, 2, \dots, M\}$) denotes the data from the ith univariate time series belonging to the m_ith modality.

In the inference phase, we aim to detect anomalies of in the multimodal time series $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_{T'}] \in \mathbb{R}^{N\times T'}$, which is from the same N univariate time series. The length of the multimodal time series is denoted as T'. The model needs to produce a detection result $\tilde{\mathbf{y}} = [\tilde{y}_1, \tilde{y}_2, \dots, \tilde{y}_{T'}] \in \mathbb{R}^{T'}$, where $\tilde{y}_i \in \{0, 1\}$ and $\tilde{y}_i = 1$ indicates the $\tilde{\mathbf{x}}_i$ is an anomaly.

3.2. Overview of MST-GAT

MST-GAT is formulated as a graph structure that treats each sensor as a node and their relationships as edges. It models complex multimodal and spatial–temporal relationships between the entire time series for MTS anomaly detection. As shown in Fig. 1, the architecture of MST-GAT involves four parts:

- **Graph Structure Learning**. It uses time series embeddings (characterizes the inherent properties of each time series) to learn a graph structure in the spatial dimension;
- **Multimodal Graph Attention Network (M-GAT)**. It captures intra- and inter-modal relations explicitly with multi-head attention module and additional relational attention modules (intra- and inter-modal attention);
- **Temporal Convolution Network**. It leverages convolutional structures on the time axis to capture temporal dependencies of time series;
- **Joint Optimization and Anomaly Score**. MST-GAT optimizes both reconstruction and prediction targets, and then identifies anomalies with the anomaly scores. The anomaly scores are further used to interpret detected anomalies.

3.3. Graph structure learning

In multimodal time series, different modalities can exhibit a variety of properties, and they are correlated in complex ways. Therefore, we prefer to use a flexible representation for each time series to capture potential correlations among multiple modalities. In this paper, we introduce time series embeddings to construct the flexible graph structure for the multi-head attention module. Consider a graph structure $\mathcal{G}$ with N nodes for multimodal time series, where each node stores the representation of a univariate time series. The edges between nodes indicate the dependency between different time series. The set of neighboring nodes of node i are denoted as $\mathcal{N}_i = \{j \mid \mathbf{A}_{ij} > 0\}$. We define the time series embedding $\mathbf{v}_i \in \mathbb{R}^d$ for node i to characterize its inherent properties, where $i \in \{1, 2, \dots, N\}$ and d is the embedding dimension. The time series embeddings are further used to construct the adjacency matrix $\mathbf{A}$ for the multi-head attention module. The adjacency matrix can be expressed as:

$$\mathbf{A}_{ij} = \mathbb{1}\quad\left\{j \in \text{TopK}\left(\left\{e_{ik} | k \in C_i\right\}\right)\right\}, \tag{4}$$

$$e_{ij} = \text{sim}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\left\|\mathbf{v}_i\right\| \times \left\|\mathbf{v}_j\right\|}, \tag{5}$$

where Eq. (4) denotes if $j \in \mathbf{C}_i$, then $\mathbf{A}_{ij} = 1$, otherwise $\mathbf{A}_{ij} = 0$, $C_i = \{1, 2, \dots, N\}$ is the candidate set, $\text{sim}(\cdot)$ is the cosine similarity, e_{ij} is the cosine similarity between node i and node j, and TopK represents the index of the largest K cosine similarity of node i selected from the candidate set. Specifically, we first compute e_{ij}, the cosine similarity between the embedding vectors. Then, we select the top K similar nodes from the candidate set to construct a sparse direct graph, and parameter K controls the sparseness of the graph structure.

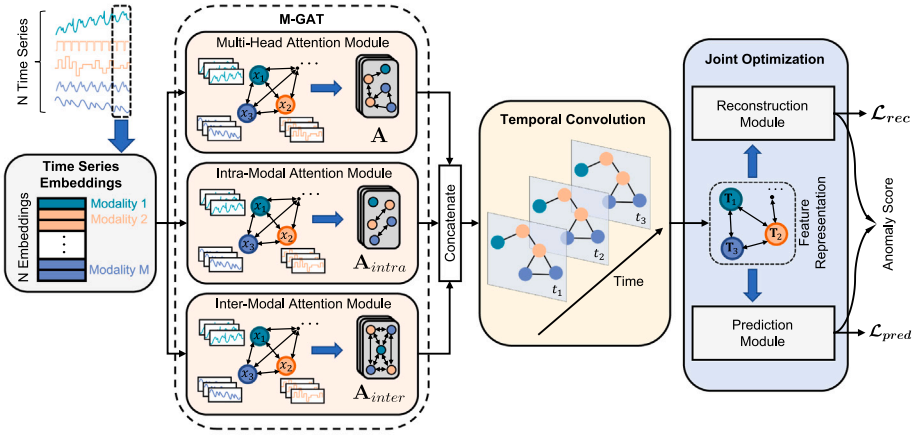

Fig. 1. MST-GAT architecture. MST-GAT consists of the M-GAT, a temporal convolution network and a joint optimization network. M-GAT captures spatial and multimodal correlations between different univariate time series. The input data is processed by M-GAT and temporal convolution to explore multimodal correlations and spatial–temporal dependencies. After that, MST-GAT uses a joint optimization network to generate reconstruction probabilities and prediction values. In the training phase, we employ the results of the joint optimization network to optimize MST-GAT. In the inference phase, the results are further used to calculate anomaly scores for anomaly detection.

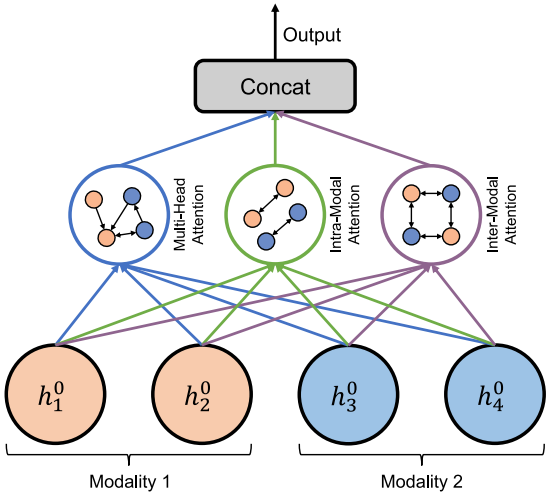

Fig. 2. The structure of the M-GAT. It includes three attention modules, i.e., multi-head attention, intra- and inter-modal attention. h_1^0, h_2^0, h_3^0 and h_4^0 represent the input features of four different univariate time series. The multi-head attention module models the modality-independent spatial relationships among multimodal time series. The intra- and inter-modal attention modules capture the multimodal correlation between different time series.

3.4. M-GAT in spatial dimension

For the training time series $\mathbf{X}$, we use the sliding window with length w to produce a fix-length input at each time. We define $\hat{\mathbf{X}}$ as the input of M-GAT at time t:

$$\hat{\mathbf{X}} = [\mathbf{x}_{t-w+1}, \mathbf{x}_{t-w+2}, \ldots, \mathbf{x}_t] \in \mathbb{R}^{N \times w}, \tag{6}$$

and we process the test time series $\tilde{\mathbf{X}}$ in the same way.

Suppose $\mathbf{H}^l$ denotes the feature representation of M-GAT at layer l. The initial input of M-GAT is $\mathbf{H}^0 = (\hat{\mathbf{X}}\mathbf{W}_{in}) \parallel \mathbf{V}$, where $\mathbf{W}_{in} \in \mathbb{R}^{w \times d}$ is the learnable transformation of input data, and $\parallel$ represents concatenation. The architecture of the proposed M-GAT is shown in Fig. 2. It consists of three attention modules, i.e., multi-head attention, intra- and inter-modal attention. The multi-head attention module focuses on modeling the modality-independent spatial relationships among multimodal time series, while intra- and inter-modal attention modules concentrate on capturing the multimodal correlation between different time series.

The multi-head attention module updates the representation of each node by aggregating the representations of its neighbors, which is formulated as:

$$h_{att_i}^{l+1} = \|_{s=1}^{S} \sum_{j \in \mathcal{N}_i} \alpha_{ij}^{ls} \mathbf{W}_{att}^{ls} h_j^l, \tag{7}$$

$$\alpha_{ij}^{ls} = \text{attention}(i, j), \tag{8}$$

where $h_j^l \in \mathbf{H}^l$ is the representation of jth node at layer l, $h_{att_i}^{l+1}$ is the feature of ith node at layer $l+1$, S denotes the number of attention heads, $\parallel$ represents concatenation, α_{ij}^{ls} represents an attention score calculated by the sth attention head at layer l between node i and node j, $\mathbf{W}_{att}^{ls}$ denotes the weight matrix of sth attention head at layer l, and $\text{attention}(i, j)$ represents the scaled dot-product attention [36].

Previous methods employ GATs to aggregate the representations of neighboring nodes according to the adjacent matrix [29]. However, these methods fail to take multimodal dependency of time series into consideration, which may lose some important information on multimodal correlations. Intuitively, neighboring nodes with different relations should have different influences on the center node. We extend the multi-head attention module with two additional relational attention modules, i.e., intra- and inter-modal attention, to integrate multimodal time series more effectively. The adjacency matrix of these two relational attention modules are defined as follows:

$$\mathbf{A}_{intra}^{ij} = \mathbb{1}\left\{j \in \mathbf{C}_{intra}^i\right\}, \tag{9}$$

$$\mathbf{A}_{inter}^{ij} = \mathbb{1}\left\{j \in \mathbf{C}_{inter}^i\right\}, \tag{10}$$

where $\mathbf{C}_{intra}^i = \left\{j | m_i = m_j\right\}$ and $\mathbf{C}_{inter}^i = \left\{j | m_i \neq m_j\right\}$ are the candidate sets. That is, $\mathbf{C}_{intra}^i$ contains nodes that belong to the same modality as node i, while $\mathbf{C}_{inter}^i$ consists of nodes that belong to different modalities from node i. In general, $\mathbf{C}_{intra}^i$ and $\mathbf{C}_{inter}^i$ are calculated by Eqs. (9) and (10), but if $|\mathbf{C}_{intra}^i| > K$, the index of the largest K cosine similarity values will be selected by TopK operation. Similarly, if $|\mathbf{C}_{inter}^i| > K$, the TopK operation will also be applied to $\mathbf{C}_{inter}^i$:

$$\mathbf{C}_{intra}^i = \left\{\text{TopK}\left(\left\{e_{ik} | k \in \mathbf{C}_{intra}^i\right\}\right)\right\}, \tag{11}$$

$$\mathbf{C}_{inter}^i = \left\{\text{TopK}\left(\left\{e_{ik} | k \in \mathbf{C}_{inter}^i\right\}\right)\right\}. \tag{12}$$

Then, we use these two relational attention modules to model the multimodal dependency among time series explicitly. We compute the feature of the intra-modal attention module as:

$$h_{intra_i}^{l+1} = \sum_{j \in \mathcal{N}_{intra_i}} \beta_{intra_i}^{lj} \mathbf{W}_{intra}^l h_j^l, \tag{13}$$

$$\beta_{intra_i}^{lj} = \frac{\exp(g_{intra_i}^{lj})}{\sum_{k\in\mathcal{N}_{intra_i}} \exp(g_{intra_i}^{lk})}, \tag{14}$$

$$g_{intra_i}^{lj} = \sigma(\mathrm{ReLU}((\mathbf{V}_i \parallel \mathbf{V}_j)\mathbf{W}_{intra1}^{l} + b_{intra1}^{l})\mathbf{W}_{intra2}^{l}), \tag{15}$$

where $h_{intra_i}^{l+1}$ is the feature of ith node at layer $l+1$, $\mathcal{N}_{intra_i} = \{j \mid \mathbf{A}_{intra}^{ij} > 0\}$ denotes the intra-modal neighbor set of node i, $\beta_{intra_i}^{lj}$ represents the attention score at layer l between node i and node j, $\mathbf{W}_{intra}^{l}$, $\mathbf{W}_{intra1}^{l}$ and $\mathbf{W}_{intra2}^{l}$ are weight matrixes at layer l, and b_{intra1}^{l} is the bias vectors at layer l. The computation of $h_{inter_i}^{l+1}$ resembles the way of computing $h_{intra_i}^{l+1}$, and $\mathcal{N}_{inter_i} = \{j \mid \mathbf{A}_{inter}^{ij} > 0\}$ is the inter-modal neighbor set of node i. We incorporate the $h_{att_i}^{l+1}$, $h_{intra_i}^{l+1}$ and $h_{inter_i}^{l+1}$ into the final representation h_i^{l+1}:

$$h_i^{l+1} = \mathrm{ReLU}(\mathbf{W}_{out}^{l+1} o_i^{l+1} + b_{out}^{l+1}), \tag{16}$$

$$o_i^{l+1} = h_{att_i}^{l+1} \parallel h_{intra_i}^{l+1} \parallel h_{inter_i}^{l+1}, \tag{17}$$

where h_i^{l+1} is the final representation of node i at layer $l+1$, $\mathbf{W}_{out}^{l+1}$ is the weight matrix at layer $l+1$, b_{out}^{l+1} is the bias vectors at layer $l+1$, $\parallel$ denotes concatenation, and o_i^{l+1} is the intermediate feature at layer $l+1$ by concatenating $h_{att_i}^{l+1}$, $h_{intra_i}^{l+1}$ and $h_{inter_i}^{l+1}$.

3.5. Convolution in temporal dimension

The multimodal graph attention captures the neighbors' information of each node in the spatial dimension, while the temporal convolution network applies the stand convolution on the time dimension to capture the temporal dynamic. The input of the temporal convolution network is the graph-level representation $\mathbf{H}^{L_{gat}}$, where L_{gat} is the number of layers in M-GAT. The temporal-level representation is calculated as:

$$\mathbf{T}^{l+1} = \mathrm{ReLU}\left(\boldsymbol{\Phi} * \left(\mathrm{ReLU}\left(\mathbf{T}^{l}\right)\right)\right), \tag{18}$$

where $\mathbf{T}^{l+1}$ denotes the temporal-level representation at layer $l+1$, $*$ is the standard convolution operation, $\boldsymbol{\Phi}$ is the kernel size, and ReLU is an activation function. The temporal convolution network updates the features of nodes by incorporating information from adjacent time slices, so it can well capture the temporal dynamics.

3.6. Joint optimization and anomaly score

The input of the reconstruction and prediction modules is the output of the temporal convolution network. For clarity, we set $\boldsymbol{\mathcal{X}}_t = \mathbf{T}^{L_{tem}}$ as the input of reconstruction and prediction modules at time t, where L_{tem} denotes the number of layers in temporal convolution. MST-GAT combines the advantages of reconstruction and prediction modules. The reconstruction module captures the data distribution of the whole time series, and the prediction module forecasts the observations at the next timestamp. We optimize MST-GAT with two tasks, i.e., reconstruction and prediction tasks. The loss function contains two optimization objectives, which are defined as:

$$\mathcal{L} = \gamma_1 \mathcal{L}_{rec} + (1-\gamma_1) \times \mathcal{L}_{pred}, \tag{19}$$

where $\mathcal{L}_{rec}$ represents the loss function for the reconstruction module, $\mathcal{L}_{pred}$ denotes the loss function for the prediction module, and γ_1 is a hyperparameter that balances the reconstruction and prediction modules.

3.6.1. Reconstruction module

The goal of the reconstruction module is to learn the reconstruction probability of the input data. Inspired by OmniAnomaly [16], we use variational autoencoder (VAE) to reconstruct $\mathcal{G}_t$. Given the input $\boldsymbol{\mathcal{X}}_t$, VAE uses the conditional distribution $p_\psi(\boldsymbol{\mathcal{X}}_t|z_t)$ to reconstruct $\mathcal{X}_t$, where z is the latent representation. The goal of training the reconstruction module is to maximize the posterior distribution of z_t:

$$p_\psi(z_t|\boldsymbol{\mathcal{X}}_t) = p_\psi(\boldsymbol{\mathcal{X}}_t|z_t)p_\psi(z_t)/p_\psi(\boldsymbol{\mathcal{X}}_t), \tag{20}$$

where $p_\psi(\boldsymbol{\mathcal{X}}_t)$ is the reconstruction probability of $\boldsymbol{\mathcal{X}}_t$. Let $p_\psi(\boldsymbol{\mathcal{X}}_t) = \{p_i|i = 1, 2, \ldots, N\}$, where p_i denotes the reconstruction probability of ith univariate time series. At each timestamp, the combined model will produce two inference results. The reconstruction probability $p_\psi(\boldsymbol{\mathcal{X}}_t)$ can be defined as follows:

$$p_\psi(\boldsymbol{\mathcal{X}}_t) = \int p_\psi(z_t)p_\psi(\boldsymbol{\mathcal{X}}_t|z_t)dz_t. \tag{21}$$

The above equation is difficult to calculate, and we need a new model $q_\rho(z_t|\boldsymbol{\mathcal{X}}_t)$ to approximate the $p_\psi(z_t|\boldsymbol{\mathcal{X}}_t)$. Given the encoder model $q_\rho(z_t|\boldsymbol{\mathcal{X}}_t)$ and the decoder model $p_\psi(\hat{\mathcal{X}}_t|z_t)$, the reconstruction loss is formulated as:

$$\begin{aligned}\mathcal{L}_{rec} = &-\mathrm{E}_{q_\rho(z_t|\boldsymbol{\mathcal{X}}_t)}[log p_\psi(\boldsymbol{\mathcal{X}}_t|z_t)] \\ &+D_{KL}(q_\rho(z_t|\boldsymbol{\mathcal{X}}_t) \parallel p_\psi(z_t)),\end{aligned} \tag{22}$$

where $\mathrm{E}_{q_\rho(z_t|\boldsymbol{\mathcal{X}}_t)}[log p_\psi(\boldsymbol{\mathcal{X}}_t|z_t)]$ denotes the log-likelihood expectation of $\boldsymbol{\mathcal{X}}_t$. D_{KL} represents the KL divergence. Negative $\mathcal{L}_{rec}$ is an estimation of the lower bound of $\log p_\psi(\boldsymbol{\mathcal{X}}_t)$.

3.6.2. Prediction module

The prediction module uses $\boldsymbol{\mathcal{X}}_t$ to predict the observations of the next timestamp. We use a multi-layer perceptron (MLP) network as the prediction module behind the temporal convolution network. The prediction loss can be defined as:

$$\mathcal{L}_{pred} = \frac{1}{T-w}\sqrt{\sum_{i=1}^{N}(x_{i,t+1} - \hat{x}_{i,t+1})^2}, \tag{23}$$

where $x_{i,t+1}$ denotes the ground truth value of ith time series at time $t+1$, and $\hat{x}_{i,t+1}$ is the forecast of ith time series at time $t+1$.

3.6.3. Anomaly score and inference

At each timestamp, the reconstruction module and the prediction module generate the reconstruction probability p_i and the forecast $\hat{\mathbf{x}}_i$, respectively, where $\hat{\mathbf{x}}_i$ denotes the prediction value of ith univariate time series. The anomaly score of MST-GAT balances the weights of these two modules. The final anomaly score of each timestamp is the sum of anomaly scores for each time series. Specifically, the anomaly score is formulated as:

$$\mathrm{score} = \sum_{i=1}^{N} \frac{(1-p_i) + \gamma_2 \times (\mathbf{x}_i - \hat{\mathbf{x}}_i)^2}{1+\gamma_2}, \tag{24}$$

where $(\mathbf{x}_i - \hat{\mathbf{x}}_i)^2$ is the square error between the forecast $\hat{\mathbf{x}}_i$ and the ground truth $\mathbf{x}_i$, and γ_2 is the hyperparameter introduced to balance the two modules, which is selected by the validation set.

In the inference phase, the detection rule is that if the anomaly score at a timestamp is greater than the defined anomaly threshold, this timestamp will be marked as "abnormal", otherwise "normal". We adopt the peaks-over-threshold (POT) algorithm [49] to select the anomaly threshold over the validation set. Finally, the overall training and inference process of MST-GAT is summarized in Algorithm 1.

4. Experiments

In this section, we conduct comprehensive experiments to demonstrate the effectiveness of MST-GAT. We first introduce the four commonly-used public datasets. Next, we evaluate MST-GAT on these datasets and show that MST-GAT performs better or on par with a range of baselines and substantially outperforms current anomaly detection methods. Then, we perform ablation studies on the key components of the proposed model. Finally, we provide the interpretability of MST-GAT through a case study.

Algorithm 1 Training and Inference Procedures of MST-GAT

Training Procedure

Input: Multimodal training time series $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$, training epochs I, batch size M and hyperparameters γ_1, γ_2.

Randomly initialize parameter W_{model} (W_{model} includes all learnable parameters in MST-GAT);
for epoch $i \in 1, 2, \ldots, I$ **do**
  Calculate $\mathbf{H}^{L_{gat}}$ in spatial dimension by M-GAT; // Eq. (16)
  Calculate $\mathbf{T}^{L_{tem}}$ in temporal dimension by temporal convolution; // Eq. (18)
  Calculate the reconstruction probability via reconstruction module; // Eq. (21)
  Calculate the prediction value via prediction module;
  Minimize the joint loss function to optimize W_{model}; // Eq. (19)
end for
return The optimized model parameter W_{model}.

Inference Procedure

Input: Multimodal testing time series $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \cdots, \tilde{\mathbf{x}}_{T'}]$, model parameter W_{model} and hyperparameters γ_1, γ_2.

for each $\tilde{\mathbf{x}}_i$ **do**
  Calculate the anomaly score of $\tilde{\mathbf{x}}_i$; // Eq. (24)
  if anomaly score > threshold **then**
    $\tilde{\mathbf{x}}_i$ = "an anomaly";
  else
    $\tilde{\mathbf{x}}_i$ = "a normal point";
  end if
end for
return Predicted label list of $\tilde{\mathbf{X}}$.

Table 1
Statistics of the four datasets used in the experiments.

Datasets	Features	Modalities	Train	Test	Anomalies (%)
MSL	27	8	58 317	73 729	10.72
SMAP	55	12	135 183	427 617	13.13
SWaT	51	8	496 800	449 919	11.98
WADI	123	8	1 048 571	172 801	5.99

4.1. Datasets

This experiment involves four benchmark datasets. We show the statistics in Table 1 and briefly introduce them in the following.

Mars Science Laboratory rover (MSL) [28] and Soil Moisture Active Passive satellite (SMAP) [28] are real-world datasets acquired from the spacecraft. These datasets are annotated by experts of NASA. Each dataset includes pre-segmented training and test sets. The training set is collected from the normal data, and the test set includes labeled anomalies. Secure Water Treatment (SWaT) [50] dataset is collected from the scaled-down water treatment testbed with 51 sensors, consisting of seven days of normal operation and four days of simulated attack scenarios. These simulated attacks include different durations and diverse attack targets. Water Distribution (WADI) [51] is a dataset acquired from a reduced water distribution testbed comprising 123 sensors. It includes two weeks under normal operation as a training set and two days with attack scenarios as a test set.

Fig. 3 shows an example of multimodal time series on WADI dataset. In the right shaded area, all sensor values have obvious fluctuations except 2_FIC_201_CO and 1_LT_001_PV, but the system is still in a normal state as these time series maintain a consistent trend. Nevertheless, in the left shaded area segment, sensor 1_AIT_001_PV behaves an inconsistent pattern compared with other univariate time series, indicating a potential problem in this sensor.

4.2. Baselines

We compare MST-GAT with eight popular MTS anomaly detection methods. They can be divided into two groups: (1) four monomodal methods, including PCA [52], AE [53], DAGMM [26], and LSTM-VAE [48]; (2) four multimodal methods, including MAD-GAN [54], OmniAnomaly [16], USAD [7], and GDN [29]. The details are presented as follows.

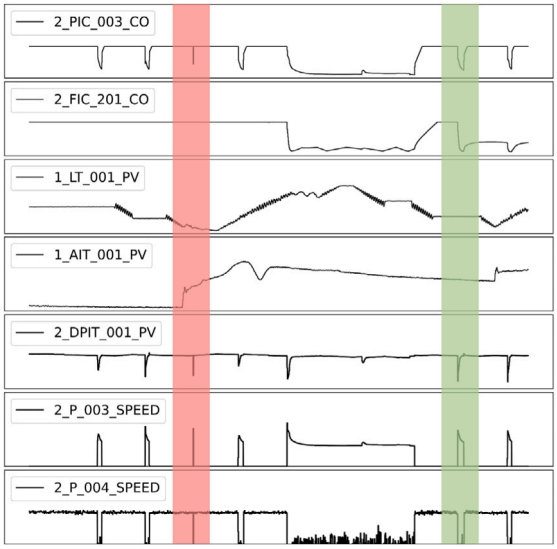

Fig. 3. An example of multimodal time series data. The left shaded area indicates anomalies, and the right shaded area indicates normal values. Time series with the same suffix belongs to the same modality (e.g., 2_P_003_$SPEED$ and 2_P_004_$SPEED$ belong to the speed modality).

- **PCA:** Principal component analysis projects the high-dimensional representation to the low-dimensional representation, and the reconstruction error of the projection is used for computing the anomaly score.
- **AE:** Autoencoder includes an encoder and a decoder and uses the reconstruction error to detect anomalies. The encoder compresses the input data into a hidden vector, and the decoder uses the vector to reconstruct the input data.
- **DAGMM:** Deep autoencoding Gaussian model combines the deep autoencoder and the Gaussian mixture model to generate the low-dimensional feature. DAGMM is a classic reconstruction-based method, which employs the reconstruction error as the anomaly score.
- **LSTM-VAE:** LSTM-VAE substitutes the fully connected network in variational autoencoder with LSTM, which can better capture the temporal dependence.
- **MAD-GAN:** Multivariate anomaly detection strategy with GAN leverages LSTM-RNN as the generator and discriminator for time series anomaly detection.
- **OmniAnomaly:** OmniAnomaly adopts a stochastic recurrent neural network for time series anomaly detection and employs the reconstruction probability to explain the detected anomalies.
- **USAD:** Unsupervised anomaly detection is an autoencoder-based framework and is trained in an adversarial fashion. The autoencoder in USAD makes the adversarial training more stable.
- **GDN:** Graph deviation network is an unsupervised anomaly detection method and uses the graph attention mechanisms to perform structure learning in multivariate time series and interprets the detected anomaly by attention weights.

4.3. Evaluation metrics for MTS anomaly detection

We use precision (Prec), recall (Rec), F1-score (F1), and the area under the ROC curve (AUC) as the evaluation metrics. The ROC curve represents a plot of true positive rate between false positive rate, and

Table 2
Comparison with existing methods in terms of precision (%), recall (%) and F1-score (%) on four datasets. The best results are highlighted in bold, and the second-best results are marked underlined. The up arrow (↑) indicates that under the Wilcoxon signed-rank test, our method achieves significant improvement compared to the baseline.

Method	MSL			SMAP			SWaT			WADI			Wilcoxon test
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	
PCA	29.37	24.14	26.50	28.84	19.93	23.57	24.92	21.63	23.16	39.53	5.63	9.86	↑
AE	71.66	50.08	58.96	72.16	79.95	75.86	72.63	52.63	61.03	34.35	34.35	34.35	↑
DAGMM	49.11	55.62	52.16	58.45	90.58	71.05	27.46	69.52	39.37	54.44	26.99	36.09	↑
LSTM-VAE	52.57	95.46	67.80	85.51	63.66	72.98	96.24	59.91	73.85	87.79	14.45	24.82	↑
MAD-GAN	85.17	89.91	87.48	80.49	82.14	81.31	98.97	63.74	77.54	41.44	33.92	37.30	↑
OmniAno.	88.67	91.17	89.90	74.16	97.76	84.34	98.25	64.97	78.22	99.47	12.98	22.96	↑
USAD	93.08	89.17	<u>91.08</u>	90.96	85.29	88.03	98.51	66.18	79.17	64.51	32.20	42.96	↑
GDN	91.35	86.12	88.66	89.32	88.72	<u>89.02</u>	99.35	68.12	<u>80.82</u>	97.50	40.19	<u>56.92</u>	↑
MST-GAT	95.06	89.10	**91.98**	91.26	89.83	**90.54**	98.73	72.41	**83.55**	98.24	43.51	**60.31**	/

AUC is defined as the area under the ROC curve. Precision is the ratio of correctly detected anomalies to all detected anomalies. Recall is the ratio of correctly detected anomalies to all actual anomalies. F1-score can comprehensively consider the value of precision and recall. For clarity, we denote true positives, false positives, and false negatives as TP, FP, and FN, respectively. Precision, recall and F1-score are formulated as:

$$\text{Prec} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{25}$$

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{26}$$

$$\text{F1} = \frac{2 \times \text{Prec} \times \text{Rec}}{\text{Prec} + \text{Rec}}. \tag{27}$$

4.4. Experimental setup

Our methods are implemented with PyTorch and trained in a Ubuntu server with Intel(R) Xeon(R) CPU E5-2640 @ 2.50 GHz and an NVIDIA 2080ti GPU. We use the Adam optimizer for training, and the learning rate is set to 1×10^{-3}. The whole network is trained with a batch size of 32 and a total of 60 epochs. The embedding dimension d is set to 128 for all datasets. We empirically set the sliding window size to 32, the kernel size of temporal convolution to 16, and the number of attention heads S to 4 for each dataset. We set the K to 15, 30, 30, and 30 for MSL, SMAP, SWaT, and WADI, respectively. The model hyperparameters γ_1 and γ_2 are selected as 0.5 and 0.8 through grid search. We utilize the POT algorithm [49] to set the anomaly threshold over the validation dataset. In the inference phase, any timestamp whose anomaly score exceeds the threshold will be considered as "abnormal".

4.5. Results and analysis

Table 2 summarizes the comparison of MST-GAT and baselines in terms of accuracy, recall and F1-score on four datasets. The results show that MST-GAT consistently outperforms the existing baselines on four benchmarks in terms of F1-score. We can observe that most baselines perform better on MSL and SMAP datasets because they have relatively simpler anomaly patterns and spatial–temporal dynamics, and MST-GAT still outperforms the best baseline by 0.9 and 1.52 in terms of F1-score (%) on MSL and SMAP datasets, respectively. Besides, most baselines show inferior results on SWaT and WADI datasets, which contain more complex anomalies, but MST-GAT significantly exceeds them by at least 2.73 in terms of F1-score (%). The structure with multimodal graph attention network and temporal convolution network efficiently integrates the information of multimodal time series, enabling MST-GAT to capture the complex spatial–temporal dynamics in multimodal time series. Not surprisingly, traditional methods (e.g., PCA and DAGMM) do not perform well compared with most deep learning methods, as they are difficult to encode comprehensive information in the time series and lack adequate consideration of the spatial and temporal dependence. Moreover, the recent USAD and GDN achieve better performance than other baselines. However, GDN is not good at obtaining temporal features from time series. USAD ignores the spatial correlation in the multimodal time series, and problems may occur when the underlying spatial dependence is complex. MST-GAT outperforms GDN that also adopts graph attention networks, showing the feasibility of using the multimodal graph attention network and temporal convolutional network. We employ the Wilcoxon signed-rank test at the 95% confidence level to identify whether the difference in performance between MST-GAT and other baselines is significant on four datasets. It is observed that our method has a significant performance improvement compared to any of the baseline methods.

We further calculate the AUC as the performance indicator on MSL and SMAP datasets, as demonstrated in Fig. 4. The proposed MST-GAT consistently outperforms other strong baselines. We attribute the performance advantage to the effective use of spatial and temporal information in multimodal time series. By using M-GAT and the temporal convolution network, MST-GAT considers the interaction of intra-modal, inter-modal and temporal information to capture the spatial–temporal correlations among multimodal time series. The results suggest that the use of multimodal graph attention network and temporal convolution network facilitates MST-GAT to achieve higher true-negative rates and lower false-positive rates in anomaly detection.

The confusion matrix of time series embeddings on MSL dataset is shown in Fig. 5. It is seen that univariate time series belonging to the same modality perform relatively high similarity, which demonstrates the intra-modal consistency on MSL dataset. As different modalities may exhibit different degrees of correlations, modality C and modality T are higher correlated with similarities in the range of $[0.29, 0.50]$, while modality P and modality T exhibit lower correlations with absolute similarities in the range of $[0.01, 0.37]$. Overall, time series embeddings strongly reflect the intra- and inter-modal correlations, revealing the feasibility of using M-GAT to model the modal dependencies between different time series.

We also conduct experiments to show the interpretability of MST-GAT. Fig. 6(a) demonstrates an anomaly example of the multimodal time series on WADI dataset. The shaded area is an anomaly interval in the multimodal time series. MST-GAT gives an average anomaly score of each sensor in the anomaly interval, and the sensor with the highest average anomaly score in the red shaded area is located as the sensor most likely to cause this anomaly. As depicted in Fig. 6(b), the darker color indicates the higher anomaly score, and $1\_AIT\_001\_PV$ is selected as the sensor that is most likely to cause this anomaly, as it shows an inconsistent trend with other time series. MST-GAT provides interpretable results that are consistent with human intuition. The ability of MST-GAT to interpret anomalies is largely attributed to the multimodal graph attention of MST-GAT, which correctly captures the correlation between features.

Furthermore, we perform sensitivity analysis of the hyperparameters on SWaT dataset. We focus on two important hyperparameters γ_1

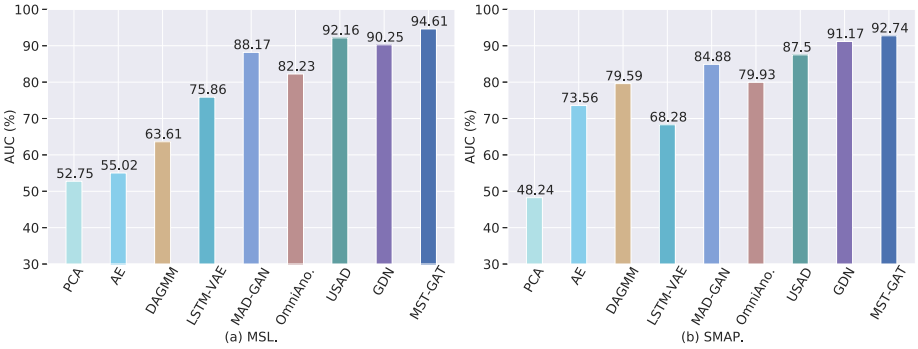

Fig. 4. AUC (%) results on MSL and SMAP datasets. The larger the better.

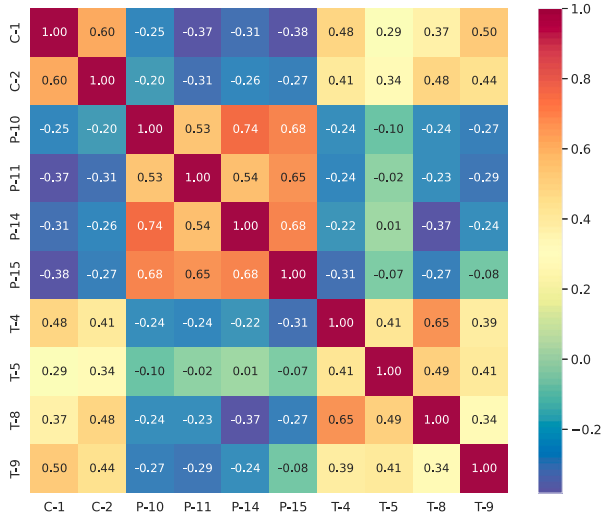

Fig. 5. Illustration of the cosine similarity of time series embeddings between three modalities that are randomly selected from MSL. The same prefix means the same modality (e.g., T-4 and T-5 belong to the temperature modality).

Table 3
Hyperparameter analysis with F1-score (%) on SWaT dataset. The best result is marked in bold.

γ_1	γ_2			
	0.4	0.6	0.8	1
0.2	82.07	82.84	83.26	82.84
0.5	82.03	83.16	**83.55**	83.45
0.8	81.97	82.50	83.35	83.01

and γ_2, which are used in the proposed loss function and anomaly score. Table 3 reports the F1-score (%) for all combinations of γ_1 from 0.2 to 0.8 with increment 0.3, and γ_2 from 0.4 to 1 with increment 0.2. The results demonstrate that MST-GAT is insensitive to γ_1 and γ_2, which exhibits robustness to different hyperparameter settings.

4.6. Ablation studies

We perform ablation experiments, which is very important to understand the role of each component of MST-GAT. We gradually remove different modules to observe the changes in performance. First, we remove the intra- and inter-modal attention modules in M-GAT. Secondly, we further remove the temporal convolution in MST-GAT. Thirdly, to study the necessity of graph structure learning, we substitute the dynamic sparse graph implemented by TopK with a complete graph. In a complete graph, all nodes are connected to each other. Finally, we discard the attention mechanism in the multi-head attention module and aggregate information by assigning equal weight to each neighbor.

Table 4
Performance comparison in terms of precision (%), recall (%), and F1-score (%) of MST-GAT and its variants. The best results are highlighted in bold.

Method	Prec	Rec	F1
MST-GAT	**98.73**	**72.41**	**83.55**
- Modal	97.36	70.12	81.52
- Temp	96.73	69.54	80.91
- TopK	91.62	65.10	76.12
- Att	70.21	67.76	68.96

The results of MST-GAT and its variants on SWaT dataset are summarized in Table 4, and we find: ***(i)*** In the experiments, removing intra- and inter-modal attention modules results in significant performance degradation, which indicates that the explicit capture of intra- and inter-modal dependencies in multimodal time series is beneficial to improving performance. We conjecture that the multimodal graph attention network is beneficial to obtain a better feature representation for MTS anomaly detection. ***(ii)*** MST-GAT equipped with temporal convolution outperforms the model without it, which shows the necessity of modeling the temporal dependence in multimodal time series. ***(iii)*** The variant of MST-GAT that does not use the attention mechanisms performs worst than the other variants. Since each univariate time series has very different properties, assigning the same weight to each neighbor introduces additional noise and cannot model the complex dependencies in multimodal time series. ***(iv)*** We can observe that the removal of each component consistently reduces performance, which proves the rationality of each component in MST-GAT.

5. Conclusions

In this paper, we have proposed a novel multimodal spatial–temporal graph attention network, termed MST-GAT, for multimodal time series anomaly detection. MST-GAT leverages the multimodal graph attention network and the temporal convolution network to capture the spatial correlation and temporal dependence among multimodal time series. The proposed model takes advantage of the reconstruction and prediction modules through joint training. Moreover, we propose an efficient anomaly interpretation approach for detected anomalies based on the reconstruction probability and the prediction value. Experimental results on benchmark datasets show that MST-GAT outperforms the state-of-the-art baselines and is able to provide interpretable results that are consistent with human intuition. In the future, it makes sense to extend our framework to unaligned multimodal time series data, e.g., multimodal data collected by sensors of self-driving cars using different sampling rates. Moreover, optimizing the memory footprint and execution time of the model to meet the needs of real-world deployments is an area worthy of future research.

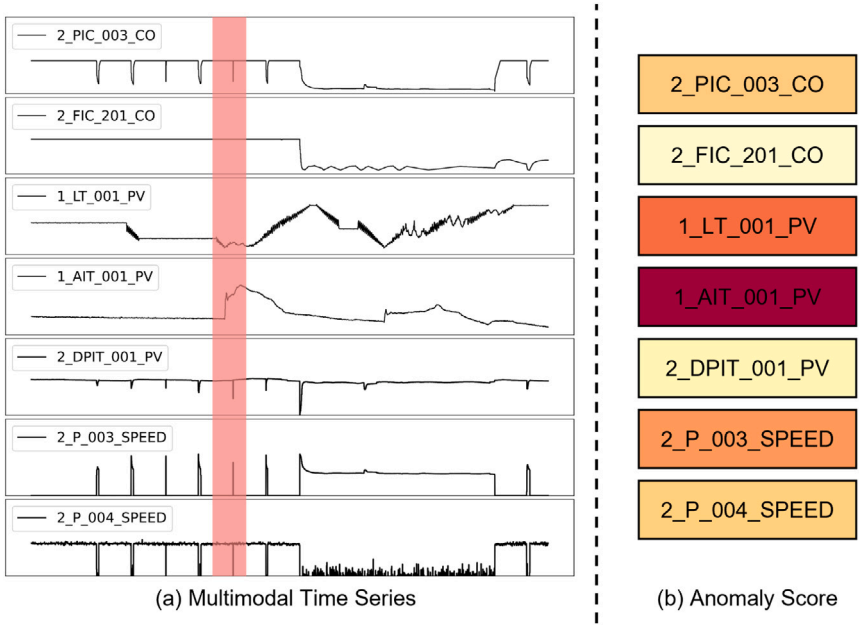

Fig. 6. The analysis of the proposed interpret method on anomalies (shaded area) of WADI dataset. Time series with the same suffix belongs to the same modality. The darker color indicates the higher anomaly score in the shaded area. The time series $1\_AIT\_001\_PV$ obtains the highest anomaly score as it shows significant inconsistencies with other time series. Therefore, $1\_AIT\_001\_PV$ is selected by MST-GAT as the sensor most likely to cause the anomaly.

CRediT authorship contribution statement

Chaoyue Ding: Conceptualization, Methodology, Software, Validation, Writing – original draft, Visualization. **Shiliang Sun:** Investigation, Conceptualization, Methodology, Supervision. **Jing Zhao:** Investigation, Conceptualization, Methodology, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgments

This work was supported by the Shanghai Municipal, Project 20511100900, the NSFC Projects 62076096 and 62006078, Shanghai Knowledge Service Platform Project ZF1213, STCSM Project 22ZR1421700 and the Fundamental Research Funds for the Central Universities.

References

[1] T. Vojir, T. Šipka, R. Aljundi, N. Chumerin, D.O. Reino, J. Matas, Road anomaly detection by partial image reconstruction with segmentation coupling, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15651–15660.

[2] L. Ruff, Y. Zemlyanskiy, R. Vandermeulen, T. Schnake, M. Kloft, Self-attentive, multi-context one-class classification for unsupervised anomaly detection on text, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4061–4071.

[3] H. Ren, B. Xu, Y. Wang, C. Yi, C. Huang, X. Kou, T. Xing, M. Yang, J. Tong, Q. Zhang, Time-series anomaly detection service at microsoft, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 3009–3017.

[4] V. Chandola, A. Banerjee, V. Kumar, Anomaly detection: A survey, ACM Comput. Surv. 41 (2009) 1–58.

[5] R. Chalapathy, S. Chawla, Deep learning for anomaly detection: A survey, 2019, arXiv preprint arXiv:1901.03407.

[6] D. Park, Z. Erickson, T. Bhattacharjee, C.C. Kemp, Multimodal execution monitoring for anomaly detection during robot manipulation, in: IEEE International Conference on Robotics and Automation, 2016, pp. 407–414.

[7] J. Audibert, P. Michiardi, F. Guyard, S. Marti, M.A. Zuluaga, USAD: Unsupervised anomaly detection on multivariate time series, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 3395–3404.

[8] Z. Li, Y. Zhao, J. Han, Y. Su, R. Jiao, X. Wen, D. Pei, Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding, in: Proceedings of the ACM SIGKDD Conference on Knowledge Discovery & Data Mining, 2021, pp. 3220–3230.

[9] R. Kromanis, P. Kripakaran, Support vector regression for anomaly detection from measurement histories, Adv. Eng. Inform. 27 (2013) 486–495.

[10] D.J. Hill, B.S. Minsker, E. Amir, Real-time Bayesian anomaly detection for environmental sensor data, in: Proceedings of the Congress-International Association for Hydraulic Research, 2007, p. 503.

[11] G.P. Zhang, Time series forecasting using a hybrid ARIMA and neural network model, Neurocomputing 50 (2003) 159–175.

[12] D.T. Shipmon, J.M. Gurevitch, P.M. Piselli, S.T. Edwards, Time series anomaly detection; detection of anomalous drops with limited features and sparse examples in noisy highly periodic data, 2017, arXiv preprint arXiv:1708.03665.

[13] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, S.Y. Philip, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. 32 (2020) 4–24.

[14] H. Zhao, Y. Wang, J. Duan, C. Huang, D. Cao, Y. Tong, B. Xu, J. Bai, J. Tong, Q. Zhang, Multivariate time-series anomaly detection via graph attention network, in: IEEE International Conference on Data Mining, 2020, pp. 841–850.

[15] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014, arXiv preprint arXiv:1412.3555.

[16] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, D. Pei, Robust anomaly detection for multivariate time series through stochastic recurrent neural network, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2828–2837.

[17] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, 2013, arXiv preprint arXiv:1312.6114.

[18] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, Graph attention networks, 2017, arXiv preprint arXiv:1710.10903.

[19] M. Braei, S. Wagner, Anomaly detection in univariate time-series: A survey on the state-of-the-art, 2020, arXiv preprint arXiv:2004.00433.

[20] L. Erhan, M. Ndubuaku, M. Di Mauro, W. Song, M. Chen, G. Fortino, O. Bagdasar, A. Liotta, Smart anomaly detection in sensor systems: A multi-perspective review, Inf. Fusion 67 (2021) 64–79.

[21] I. Kiss, B. Genge, P. Haller, G. Sebestyén, Data clustering-based anomaly detection in industrial control systems, in: IEEE International Conference on Intelligent Computer Communication and Processing, 2014, pp. 275–281.

[22] W.A. Chaovalitwongse, Y.-J. Fan, R.C. Sachdeo, On the time series k-nearest neighbor classification of abnormal brain activity, IEEE Trans. Syst. Man Cybern. A 37 (2007) 1005–1016.
[23] J. Ma, S. Perkins, Time-series novelty detection using one-class support vector machines, in: Proceedings of the International Joint Conference on Neural Networks, 2003, pp. 1741–1745.
[24] L. Puggini, S. McLoone, An enhanced variable selection and isolation forest based methodology for anomaly detection with OES data, Eng. Appl. Artif. Intell. 67 (2018) 126–135.
[25] C. Zhang, D. Song, Y. Chen, X. Feng, C. Lumezanu, W. Cheng, J. Ni, B. Zong, H. Chen, N.V. Chawla, A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 1409–1416.
[26] B. Zong, Q. Song, M.R. Min, W. Cheng, C. Lumezanu, D. Cho, H. Chen, Deep autoencoding Gaussian mixture model for unsupervised anomaly detection, in: International Conference on Learning Representations, 2018, pp. 1–14.
[27] L. Shen, Z. Yu, Q. Ma, J.T. Kwok, Time series anomaly detection with multiresolution ensemble decoding, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 9567–9575.
[28] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, T. Soderstrom, Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 387–395.
[29] A. Deng, B. Hooi, Graph neural network-based anomaly detection in multivariate time series, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2021, pp. 4027–4035.
[30] W. Yuan, K. He, D. Guan, L. Zhou, C. Li, Graph kernel based link prediction for signed social networks, Inf. Fusion 46 (2019) 1–10.
[31] S.-H. Wang, V.V. Govindaraj, J.M. Górriz, X. Zhang, Y.-D. Zhang, Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network, Inf. Fusion 67 (2021) 208–229.
[32] T.N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, 2016, arXiv preprint arXiv:1609.02907.
[33] D. Duvenaud, D. Maclaurin, J. Aguilera-Iparraguirre, R. Gómez-Bombarelli, T. Hirzel, A. Aspuru-Guzik, R.P. Adams, Convolutional networks on graphs for learning molecular fingerprints, Adv. Neural Inf. Process. Syst. (2015) 2224–2232.
[34] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, M.M. Bronstein, Geometric deep learning on graphs and manifolds using mixture model CNNs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5115–5124.
[35] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, 2014, arXiv preprint arXiv:1409.0473.
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[37] X. Zhang, C. Huang, Y. Xu, L. Xia, Spatial-temporal convolutional graph attention networks for citywide traffic flow forecasting, in: Proceedings of the ACM International Conference on Information & Knowledge Management, 2020, pp. 1853–1862.
[38] R.-G. Cirstea, C. Guo, B. Yang, Graph attention recurrent neural networks for correlated time series forecasting–full version, 2021, arXiv preprint arXiv:2103.10760.
[39] L. Zhu, B. Wan, C. Li, G. Tian, Y. Hou, K. Yuan, Dyadic relational graph convolutional networks for skeleton-based human interaction recognition, Pattern Recognit. 115 (2021) 107920.
[40] S. Poria, E. Cambria, R. Bajpai, A. Hussain, A review of affective computing: From unimodal analysis to multimodal fusion, Inf. Fusion 37 (2017) 98–125.
[41] X. Liu, J. Zhao, S. Sun, H. Liu, H. Yang, Variational multimodal machine translation with underlying semantic alignment, Inf. Fusion 69 (2021) 73–80.
[42] H. Wen, Y. Liu, I. Rekik, S. Wang, Z. Chen, J. Zhang, Y. Zhang, Y. Peng, H. He, Multi-modal multiple kernel learning for accurate identification of Tourette syndrome children, Pattern Recognit. 63 (2017) 601–611.
[43] Y. Zhen, D.-Y. Yeung, A probabilistic model for multimodal hash function learning, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2012, pp. 940–948.
[44] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, J. Huang, Deep multimodal fusion by channel exchanging, in: Advances in Neural Information Processing Systems, 2020, pp. 4835–4845.
[45] B.K. Iwana, S. Uchida, Time series classification using local distance-based features in multi-modal fusion networks, Pattern Recognit. 97 (2020) 107024.
[46] P. Yang, B. Chen, P. Zhang, X. Sun, Visual agreement regularized training for multi-modal machine translation, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 9418–9425.
[47] S. Nedelkoski, J. Cardoso, O. Kao, Anomaly detection from system tracing data using multimodal deep learning, in: IEEE International Conference on Cloud Computing, 2019, pp. 179–186.
[48] D. Park, Y. Hoshi, C.C. Kemp, A multimodal anomaly detector for robot-assisted feeding using an LSTM-based variational autoencoder, IEEE Robot. Autom. Lett. 3 (2018) 1544–1551.
[49] A. Siffer, P.-A. Fouque, A. Termier, C. Largouet, Anomaly detection in streams with extreme value theory, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2017, pp. 1067–1075.
[50] J. Goh, S. Adepu, K.N. Junejo, A. Mathur, A dataset to support research in the design of secure water treatment systems, in: International Conference on Critical Information Infrastructures Security, 2016, pp. 88–99.
[51] C.M. Ahmed, V.R. Palleti, A.P. Mathur, WADI: a water distribution testbed for research in the design of secure cyber physical systems, in: Proceedings of the International Workshop on Cyber-Physical Systems for Smart Water Networks, 2017, pp. 25–28.
[52] M.-L. Shyu, S.-C. Chen, K. Sarinnapakorn, L. Chang, A novel anomaly detection scheme based on principal component classifier, in: Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, 2003, pp. 172–179.
[53] O.I. Provotar, Y.M. Linder, M.M. Veres, Unsupervised anomaly detection in time series using lstm-based autoencoders, in: IEEE International Conference on Advanced Trends in Information Theory, 2019, pp. 513–517.
[54] D. Li, D. Chen, B. Jin, L. Shi, J. Goh, S.-K. Ng, MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks, in: International Conference on Artificial Neural Networks, 2019, pp. 703–716.

Chaoyue Ding is currently pursuing the M.S. degree with the Pattern Recognition and Machine Learning Research Group, School of Computer Science and Technology, East China Normal University, Shanghai, China. His current research interests include anomaly detection and time series modeling.

Shiliang Sun is a professor with the School of Computer Science and Technology and the head of the Pattern Recognition and Machine Learning Research Group, East China Normal University, Shanghai, China. He received the Ph.D. degree in pattern recognition and intelligent systems from the Department of Automation and the State Key Laboratory of Intelligent Technology and Systems, Tsinghua University, Beijing, China, in 2007. From 2009 to 2010, he was a visiting researcher with the Department of Computer Science, University College London, U.K. In 2014, he was a visiting researcher with the Department of Electrical Engineering, Columbia University, New York. His current research interests include kernel methods, multi-view learning, learning theory, approximate inference, sequential modeling, and their applications. His research results have expounded in 200+ publications at peer-reviewed journals and conferences, such as the Journal of Machine Learning Research, the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Transactions on Cybernetics, NIPS, ICML, IJCAI, AAAI. He is on the editorial board of multiple international journals, including the Information Fusion and the IEEE Transactions on Neural Networks and Learning Systems.

Jing Zhao is an associate professor in the Pattern Recognition and Machine Learning Research Group, School of Computer Science and Technology, East China Normal University. Her research interests include Bayesian methods, sequence modeling, deep learning and their applications. Her research results have expounded in 30+ publications at peer-reviewed journals and conferences, such as the Journal of Machine Learning Research, the IEEE Transactions on Neural Networks and Learning Systems, the IEEE Transactions on Cybernetics, and IJCAI.