**Journal of Computer Technology and Software**

**ISSN:2998-2383**

**Vol. 4, No. 9, 2025**

# Structure-Guided Attention and Multi-Scale Behavior Modeling for Microservice Anomaly Detection

**Xuexian Li[1], Ruohan Yin[2], Siyuan Dang[3]**
[1]Northeastern University, Boston, USA
[2]University of New Hampshire, Durham, USA
[3]Stevens Institute of Technology, New Jersey, USA
*Corresponding author: Siyuan Dang; sheboke93@gmail.com

**Abstract:** This paper proposes an anomaly detection algorithm based on an improved Transformer architecture to address the limitations in modeling complex behavioral dependencies and capturing asynchronous service anomalies in microservice systems. The method integrates graph structure awareness and multi-scale behavior modeling. A structure-guided attention module is introduced to enhance the accuracy of modeling topological dependencies in service invocation graphs. In addition,multi-scale convolution and residual paths are used to build a hybrid representation space for short-term and long-term service behaviors, improving the model's sensitivity to sparse and burst anomalies. The overall architecture consists of a graph representation learning layer, a multi-head global attention layer, and a structure-level fine-tuning module, enabling layered feature abstraction and anomaly distribution representation. Experiments are conducted on real-world microservice datasets. Results show that the proposed method outperforms mainstream baseline models in F1-score, AUROC, and AUPR, confirming its applicability, stability, and detection accuracy in complex microservice environments. The proposed framework demonstrates strong robustness, scalability, and structural generalization, offering an effective modeling paradigm for anomaly perception tasks in high-dimensional and heterogeneous microservice systems.

**Keywords:** Structure-guided attention; microservice system; anomaly modeling framework; multi-scale feature fusion

## 1. Introduction

With the rapid development of cloud-native architectures and container orchestration technologies, microservices have become the dominant paradigm for building modern distributed systems. By decomposing monolithic applications into multiple independent services, microservice architecture significantly improves system flexibility, scalability, and modularity. However, this increased service granularity also leads to a highly complex runtime environment. Invocation chains among services have become more dynamic and asynchronous. As a result, system observability and interpretability have declined. In scenarios involving multi-tenancy,multi-protocol interactions, and multi-dimensional metrics, traditional monitoring methods struggle to accurately capture the evolution of potential anomalies [1].

Microservice anomaly detection, as a core technology for ensuring system stability, faces multiple challenges. First, anomalous behaviors often manifest as short-term, sparse bursts or cross-service propagation paths. Traditional methods based on static rules or statistical thresholds fail to capture such complex nonlinear variations. Second, anomalies frequently involve implicit changes in service dependencies and dynamic adjustments in behavioral context, making it difficult to achieve accurate detection using only local features. In real-world production environments, how to balance detection accuracy with real-time performance and system scalability has become a key issue in model design.

In recent years, deep learning-based methods have made significant progress in complex sequence modeling tasks. The Transformer architecture, in particular, has demonstrated strong modeling capabilities in fields such as natural language processing and time series forecasting, due to its global modeling and self-attention mechanisms. However, the standard Transformer still has notable limitations when applied to microservice anomaly detection. These include a lack of structural awareness, poor capability in modeling sparse behavioral dependencies, and performance degradation in asynchronous service interaction scenarios. Therefore, relying solely on the standard Transformer is insufficient for handling the complexity of heterogeneous topology and dynamic behavior in microservice systems.

To address these issues, there is an urgent need to build an anomaly detection architecture that combines structure-guided modeling with the ability to capture multi-scale behavioral evolution. Such an architecture should integrate service invocation topology, contextual behavior sequences, and anomaly signal features into a unified representation space. This would enable accurate localization and identification of potential anomaly paths. In the context of rapidly evolving microservice systems and increasing instability in runtime states, developing new models with long-term dependency modeling and structure-sensitive feature extraction is not only crucial for enhancing system robustness but also a key step in building intelligent operation and maintenance systems [2].

## 2. Related Work

Microservice anomaly detection has long been a focal point in the field of intelligent system operations. Early approaches largely relied on static rule bases and predefined statistical thresholds. The core idea was to extract key indicators based on expert knowledge and define fixed alarm rules. These methods provided high interpretability and fast response in monolithic or low-complexity environments. However, they lack adaptability and are costly to update when applied to large-scale, dynamic, and heterogeneous microservice architectures. In addition, traditional statistical modeling methods such as ARIMA and EWMA can capture trends and periodic patterns from time series. Yet, they tend to lag when faced with nonlinear fluctuations or abrupt changes in anomalous behaviors, making them insufficient for real-time monitoring of complex invocation chains.

With the rapid advancement of machine learning, supervised and semi-supervised detection approaches have gradually become mainstream. These methods rely on historical data for training and use classifiers or regressors to determine system states. Classical models such as support vector machines, decision trees, and random forests have shown effectiveness in some cases due to their structural representation capacity. However, they are often sensitive to feature dependencies and heavily reliant on labeled data. In practical deployment, their effectiveness is constrained by annotation costs and sample distribution shifts. Furthermore, the application of deep learning to unsupervised anomaly detection has introduced new opportunities for modeling complex temporal patterns and high-dimensional behaviors. Methods such as autoencoders, variational inference models, and temporal convolutional networks demonstrate strong capabilities in feature abstraction and anomaly characterization under multimodal inputs. Nonetheless, they still struggle to model long-range dependencies and context shifts across components accurately [3].

Among deep architectures, the Transformer has demonstrated remarkable advantages in complex sequence modeling tasks, due to its parallel computation capability and multi-head attention mechanism. Its key strength lies in capturing dynamic dependencies between arbitrary positions, which makes it suitable for asynchronous calls and long-distance relationships in microservices. However, the original Transformer was designed for natural language tasks. When applied to microservice time series data, it faces several challenges. First, service behavior data is often sparse, heterogeneous, and high-dimensional, which introduces significant noise during attention computation. Second, the Transformer cannot model structural changes in sequences, making it difficult to detect subtle anomalies. In addition, its dependence on fixed positional encoding weakens its robustness in real-world scenarios where service execution orders change frequently.

To address these challenges, recent studies have explored structural enhancements to improve Transformer adaptability in system anomaly detection tasks. Some approaches integrate convolutional modules to enhance local pattern modeling. Others incorporate graph neural networks to represent topological dependencies among components. Sparse attention mechanisms have also been used to reduce redundant computation in large-scale invocation graphs. These strategies mitigate certain performance limitations of the Transformer in practical deployment. However, they do not yet offer a systematic solution to the core modeling challenges related to asynchrony, sparsity, and multi-scale temporal dynamics in service behaviors. Therefore, a critical technical path in current research is to develop structural-level enhancements that retain the global modeling strength of the Transformer while building an efficient, scalable, and generalizable modeling framework tailored for microservice anomaly detection [4].

## 3. Method

This paper proposes a network architecture that centers on a modified Transformer and integrates graph representation learning with a structural segmentation fine-tuning module. The framework enables precise modeling of multi-scale dependencies and contextual anomalies in microservice behavior sequences. In the frontend, global structure-aware representations are constructed through multi-client private data training and graph embedding extraction. In the backend, a structure-level fine-tuning mechanism is introduced to adapt to dynamic coupling relationships between services. The overall design achieves robust and highly sensitive anomaly detection performance. The model architecture is shown in Figure 1.

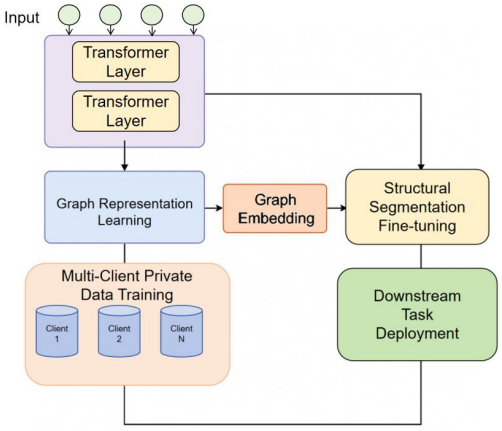


**Figure 1.** Improved Transformer-based Microservice Anomaly Detection Framework

This study proposes an improved Transformer architecture that integrates a structural enhancement mechanism and multi-scale modeling strategy to improve the anomaly detection capability in microservice behavior sequences. This method takes multi-source service behavior logs as input, constructs a unified embedding representation, and captures dynamic behavior dependencies at different scales through local and global path modeling. Specifically, the model first encodes each time step in the service call sequence, extracts high-dimensional feature representations, and establishes an embedding matrix for the input sequence. Let the input sequence be $\{x_1, x_2, ..., x_T\}$, and its corresponding

embedding representation be $\{h_1, h_2, ..., h_T\}$ . The initial embedding can be defined by the following mapping function:

$$h_t = Embed(x_t) + PE(t)$$

Among them, $PE(t)$ represents the position encoding term represents, which is combined with explicit time order information to enhance the model's perception of the behavior sequence structure. To effectively characterize the long-range dependency structure, a multi-head attention mechanism is introduced to construct the contextual interaction weights between services and then generate an aggregate vector that represents the global behavior dependency. The calculation method of each attention head is as follows:

$$Attention(Q, K, V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V$$

Among them, $Q, K, V$ represents the query, key, and value vectors represent,respectively, and $d_k$ is the feature dimension. This mechanism can dynamically adjust the attention weights between different positions to capture the non-local dependencies between services and avoid losing key abnormal signals due to local window limitations. At the same time, to make up for the problem that the standard Transformer is not sensitive enough to micro-perturbations, a structural attention routing mechanism is designed to guide attention to flow in the structural graph before emphasizing the key transfer edges in the behavior path. Assuming that the topological structure of the service call is represented by the adjacency matrix $A$ , then the structure-guided attention matrix can be expressed as:

$$\widetilde{A} = Normalize(A \otimes Attention(Q, K, V))$$

Among them, $\otimes$ represents the element-level product, and $\widetilde{A}$ retains the dependency flow under the structural constraints, which improves the model's robustness in modeling the abnormal propagation chain. Furthermore, to enhance the model's ability to identify local behavior fluctuations and short-term anomalies, this paper introduces a multi-scale convolution path to extract the embedded sequence in parallel, and adopts a residual mechanism in the fusion stage to retain the original dynamic features. Multi-scale convolution can be formalized as:

$$c_t^{(k)} = \sigma(W^{(k)} * h_{t:t+k-1} + b^{(k)})$$

Among them, $k$ represents the convolution kernel size, $\sigma$ is the activation function, $W^{(k)}$ and $b^{(k)}$ are the convolution weight and bias terms, respectively, and local dynamic patterns are extracted through convolution kernels of different scales. At the output end, the model integrates multi-head global representation and multi-scale local features, and maps them into the final abnormal characterization result through a unified decoding network. The final representation output is:

$$z_t = Fusion(h_t^{global}, c_t^{multi-scale})$$

The output vector $z_t$ combines long-term and short-term behavior dependencies with structural path information and can comprehensively model potential anomalies in microservice behavior sequences. The overall method introduces structural guidance and scale fusion mechanisms while maintaining the global modeling advantages of the Transformer, effectively improving the model's anomaly recognition and expression capabilities in complex scenarios.

## 4. Experimental Results

### 4.1 Dataset

This article uses the "Microservice Bottleneck Location Dataset" on Kaggle. The dataset is collected from a microservice-based social network application deployed on Kubernetes using the DeathStarBench microservice suite. It contains call logs, performance metrics, and bottleneck annotations under different load conditions. The dataset was released after 2021 and is available through the Kaggle search. Its characteristics are closely related to multi-service call chains and asynchronous behavior, which is very suitable for microservice anomaly detection tasks.

The dataset contains key metrics such as inter-service call latency, TPS (transactions per second), CPU utilization, and memory consumption. It also provides labels for bottleneck services, which helps to build objective functions and evaluate the detection performance of structure-sensitive models. The data structure combines service call graphs with performance event records, providing rich temporal and topological information for research tasks.

Applying this dataset to our research helps to verify the proposed improved Transformer model in a real microservice environment. It can evaluate the model's ability to identify performance anomalies and locate bottlenecks. This is consistent with the design goals of the model, namely capturing multi-scale dependencies, implementing structure-guided attention mechanisms, and improving anomaly sensitivity.

### 4.2 Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

The experimental results show that the proposed model based on the improved Transformer architecture (IT-MAD) achieves superior performance across all evaluation metrics. It especially outperforms existing methods in terms of F1-score and AUROC, demonstrating high robustness and sensitivity in microservice behavior modeling. By introducing structure-aware mechanisms and multi-scale feature construction paths, the model enhances the ability to identify anomaly propagation paths and cross-component dependencies. It effectively mitigates the modeling degradation problems faced by traditional Transformers in asynchronous service interaction scenarios.

**Table1.** Comparative experimental results

| Model | Precision | Recall | F1-score | AUROC | AUPR |
|---|---|---|---|---|---|
| Ours (IT-MAD) | 0.837 | 0.789 | 0.812 | 0.871 | 0.846 |
| DeepSAD [5] | 0.771 | 0.728 | 0.749 | 0.823 | 0.795 |
| TranAD [6] | 0.788 | 0.742 | 0.764 | 0.842 | 0.814 |
| RADAR [7] | 0.801 | 0.765 | 0.782 | 0.857 | 0.829 |
| GDN [8] | 0.775 | 0.749 | 0.761 | 0.835 | 0.811 |

Compared to baseline models such as TranAD and GDN, IT-MAD maintains global modeling capability while incorporating graph embeddings and structure-guided attention fusion. This allows the model to capture complex invocation topologies and behavioral state evolution in microservice systems. The fusion mechanism improves the model ' s capacity to represent higher-order dependencies and increases its responsiveness to subtle perturbations in behavior sequences. As a result, the model becomes more stable in identifying low-frequency and high-impact anomalies.

Furthermore, the improvement in the AUPR metric indicates that the proposed model performs better under sparse anomaly distributions. This is especially important for detecting highly imbalanced data distributions in real-world microservice systems. Through structure enhancement and joint temporal-topological modeling, IT-MAD effectively reduces false positives and false negatives. It maintains high recall while ensuring high precision, thereby improving overall detection performance.

Overall, the model demonstrates better comprehensive performance than existing mainstream approaches. It verifies the adaptability and generalization potential of the improved Transformer in microservice anomaly detection tasks. The design extends from sequence-level modeling to structure-level modeling. It enables deep modeling of dynamic service behaviors while keeping computational costs low, providing both theoretical support and empirical evidence for building more intelligent and stable service operation systems.

This paper also experiments on the impact of different numbers of Transformer encoding layers on anomaly detection performance. The experimental results are shown in Figure 2.

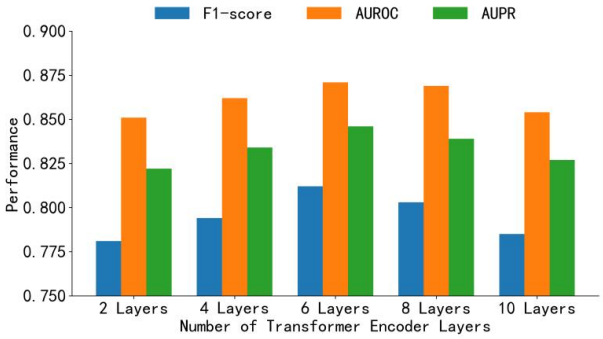


**Figure 2.** The impact of different Transformer encoding layers on anomaly detection performance

As the number of Transformer encoder layers increases, the model's anomaly detection performance shows an overall trend of first improving and then declining. The best results are achieved with a six-layer encoder structure. This suggests that an appropriate depth helps the model capture complex dependencies and asynchronous patterns in microservice behavior sequences. It balances structure enhancement and contextual modeling, thereby improving global awareness and local anomaly sensitivity.

When the number of layers is low, such as two or four, the model's representation capacity is limited. It struggles to establish deep behavioral dependencies across services, leading to lower F1-score and AUPR. This indicates that shallow structures are insufficient for extracting key information in microservice scenarios. The problem is more pronounced when service invocation chains are long or anomalies exhibit structural coupling characteristics.

As the depth increases to eight and ten layers, model performance degrades to some extent. This reflects that excessive depth may introduce redundant computation and attention diffusion. It weakens the model's ability to capture fine-grained anomaly boundaries. In particular, during multi-scale information fusion, too many layers can cause representation shift and saturation, reducing the sensitivity to low-frequency burst anomalies.

These results confirm that the number of encoder layers is a critical structural hyperparameter in microservice anomaly detection. Its selection significantly affects modeling quality. Proper layer design enhances the modeling capacity of the improved Transformer. It also controls model complexity while preserving structural expressiveness, further improving the stability and generalizability of the model in practical deployment.

This paper also conducted a comparative experiment on the regulatory effect of structure-guided attention intensity parameters on detection performance. The experimental results are shown in Figure 3.

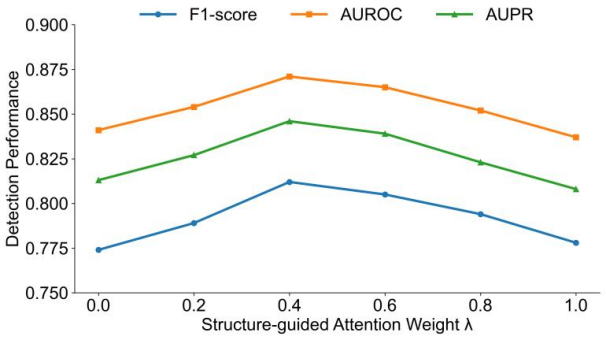


**Figure 3.** Structure-guided attention intensity parameters modulate detection performance

Experimental results show that as the structural-guided attention strength parameter λ increases from 0, the overall detection performance of the model first improves and then declines. This indicates that moderately introducing structural information can significantly enhance the performance of the improved Transformer in microservice anomaly detection. When λ is set to 0.4, the model achieves the best F1-score, AUROC, and AUPR. This suggests that the structure-guided mechanism effectively incorporates topological constraints

between services while preserving the model's ability to capture behavioral context, thereby improving the structural recognition of anomaly patterns.

When λ is set to a low value, such as 0.0 or 0.2, the model relies primarily on the original self-attention mechanism for information integration. It fails to fully leverage the inherent dependency paths among microservices. As a result, the model struggles to detect cross-service propagated anomalies. Under such settings, its ability to perceive anomaly signals in complex service graphs is limited, especially in scenarios with long invocation paths or strong coupling intensity.

As λ increases further to 0.6 or even 1.0, model performance declines. This may be due to the structural information dominating the attention mechanism, which suppresses the modeling of temporal dynamics in behavioral sequences. Excessive dependence on structure-guided attention may cause attention weights to concentrate too heavily on static paths in the graph. This reduces the model's sensitivity to behavioral variations and concurrent anomalies, which are common in microservice systems.

The experiment confirms that the structural-guided attention weight has a significant regulatory effect on model performance. An appropriate λ value balances the relationship between structure awareness and behavior modeling. It preserves the global attention mechanism's advantage in representing long-range behavioral dependencies while enhancing the model's ability to respond to topological anomaly propagation. This supports more effective identification of complex anomaly patterns in microservice systems.

## 5. Conclusion

This paper addresses the core challenges of anomaly detection in microservice systems and proposes an improved detection framework that integrates graph structure modeling with a multi-scale Transformer mechanism. By introducing structure-guided attention and multi-scale dynamic modeling paths, the method effectively mitigates the limitations of traditional models in handling heterogeneous invocation chains and cross-component behavioral dependencies. Experimental results demonstrate that the proposed model outperforms mainstream methods across multiple key performance metrics, confirming the effectiveness and adaptability of the structure-aware and temporal modeling fusion strategy in complex microservice environments.

Specifically, the graph representation learning module designed in the frontend significantly enhances the system’s ability to identify topological relationships among services. This reduces the risk of missing or misclassifying unstructured anomalies. The structure segmentation fine-tuning mechanism in the backend improves sensitivity to sparse and low-frequency burst anomalies. The overall model architecture offers flexibility in modeling cross-service dependencies while maintaining computational efficiency and deployment practicality. It provides both theoretical support and engineering foundations for real-time anomaly perception in large-scale microservice systems.

From an application perspective, the proposed framework shows strong generality and scalability. It can be applied to various complex service infrastructures, including cloud-native platforms, container orchestration systems, and edge computing networks. In areas such as system security, automated operations, and elastic resource scheduling, the method can be integrated into existing monitoring architectures to ensure business stability and service availability. Furthermore, the deep coupling of structural information and temporal behavior provides a unified modeling foundation for extended tasks such as service dependency analysis and root cause localization.

Future research can expand the capabilities of this framework from multiple dimensions. One possible direction is to incorporate causal graph modeling to enhance the ability to trace anomaly sources. Another is to adapt the method to federated learning environments for multi-source heterogeneous microservice architectures, enabling collaborative anomaly detection with privacy protection. As the integration of large language models and multimodal system operation data becomes increasingly important, extending this method to support cross-modal information fusion and semantically enhanced anomaly modeling will be a promising direction for further investigation..

## References


[1] L. Akmeemana, C. Attanayake, H. Faiz et al., "GAL-MAD: Towards Explainable Anomaly Detection in Microservice Applications Using Graph Attention Networks," arXiv preprint arXiv:2504.00058, 2025.

[2] Liu, Z., Meng, R., Huang, S. Y., & Huang, Z. "Cost-Sensitive Mamba Sequence Modeling for Fault Detection in Cloud-Native Microservice Systems", 2025.

[3] C. Ding, S. Sun and J. Zhao, "MST-GAT: A Multimodal Spatial–Temporal Graph Attention Network for Time Series Anomaly Detection," Information Fusion, vol. 89, pp. 527-536, 2023.

[4] Z. Li, J. Zhao and J. Kang, "Multi-source Anomaly Detection for Microservice Systems," Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings, pp. 414-415, 2024.

[5] L. Ruff, R. A. Vandermeulen, N. Görnitz et al., "Deep Semi-Supervised Anomaly Detection," arXiv preprint arXiv:1906.02694, 2019.

[6] S. Tuli, G. Casale and N. R. Jennings, "TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data," arXiv preprint arXiv:2201.07284, 2022.

[7] Y. Wang, C. Qin, Y. Bai et al., "Making Reconstruction-Based Method Great Again for Video Anomaly Detection," Proceedings of the 2022 IEEE International Conference on Data Mining (ICDM), pp. 1215-1220, 2022.

[8] S. Ray, S. Lakdawala, M. Goswami et al., "Learning Graph Neural Networks for Multivariate Time Series Anomaly Detection," arXiv preprint arXiv:2111.08082, 2021.