RESTAD: Residual-Enhanced Spatio-Temporal Modeling with Multi-Scale Attention for Multivariate KPI Anomaly Detection

Shuyi Shang[1,2,*]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[*]Corresponding author: Shuyi Shang E-mail: shangshuyi@iie.ac.cn

***Abstract*—Multivariate Key Performance Indicators (KPIs) are essential time-series data for characterizing the real-time operational status of Internet-based services, where precise anomaly detection is fundamental to maintaining system stability. Although current unsupervised reconstruction models learn normal patterns by mining spatio-temporal dependencies, they struggle to capture fine-grained features amidst noise interference. Addressing the limitation where current models bias toward fitting low-frequency trends while neglecting high-frequency fluctuations, we observe that the high-frequency components, which are easily relegated to the residual space by base learners, actually carry critical spatio-temporal correlation patterns that reflect system state transitions. Specifically, a base learner leveraging spatial attention mechanism is first employed to filter out residuals. Subsequently, a dual-scale temporal attention learner models both low-frequency trends and short-term fine-grained features within the residuals. Furthermore, we introduce a memory-enhanced cross-attention mechanism to refine residual representations. Simultaneously, a spatial consistency constraint based on KL divergence is incorporated to strengthen the modeling of topological associations among multivariate KPIs. Extensive experimental results demonstrate that RESTAD outperforms state-of-the-art methods across four real-world datasets, exhibiting a superior capability to identify anomalies from intricate high-frequency backgrounds.**

***Index Terms*—time series analysis, anomaly detection, residual-enhanced learning, attention mechanism**

I. Introduction

With the rapid popularization of the Internet, Internet-based services have permeated virtually every industry and sector in modern society, including financial transactions, industrial remote monitoring, medical diagnosis, and multimedia services. Within these services, any minor anomalous events may trigger service quality degradation or even disruption. Consequently, the Artificial Intelligence for IT Operations (AIOps) supporting these services is critical. As time series data that effectively capture the real-time operational status of Internet-based services and their underlying hardware, key performance indicators (KPIs) have been confirmed to reduce storage and communication costs by more than 85% for anomaly detection tasks in AIOps[1].

Although supervised models[2] have demonstrated remarkable performance in KPI anomaly detection across prior scenarios, manual labeling of millions of KPI data points poses a significant challenge amid the growing scale and complexity of such services. Consequently, unsupervised approaches have been widely applied in practice. KPI anomalies are typically defined as temporal instances that exhibit significant deviations from normal behaviors. Classic machine learning methods, including isolation forest models [3] and density estimation-based approaches [4], generally extract features along the time dimension and then identify anomalies from the extracted discrete features. However, these approaches overly rely on the quality of feature extraction, rendering them challenging to generalize across diverse application scenarios.

Deep learning methods, particularly reconstruction-based models [5], have provided robust capabilities for multivariate KPI anomaly detection by capturing intricate spatio-temporal dependencies. While existing works model static topologies [1] or capture dynamic spatio-temporal dependencies within localized segments, they tend to be dominated by low-frequency periodic patterns to maintain training stability against noise. This bias inadvertently neglects high-frequency fluctuations prevalent in real-world KPIs. We find that these instantaneous high-frequency spikes, rather than being mere stochastic noise, are often the primary carriers of critical latent spatio-temporal correlations.

To address these issues, we propose RESTAD, a residual-enhanced spatio-temporal model for multivariate KPI anomaly detection, which proactively leverages residuals derived from a foundational learner. Technically, RESTAD employs a global spatial-attention base learner to model inter-metric correlations within the raw KPIs and calculate the residuals. By capturing dominant spatio-aware structural representations, the base learner leaves high-frequency fluctuations in the residual space as unexplained components. Then, we adopt a patch-wise temporal attention learner to recover dynamic local spatial correlations hidden within high-frequency fluctuations. Inspired by[1], to mitigate the computational overhead of long sequences, we leverage a remote observation window to replace all preceding remote accumulated residuals, capturing low-frequency trends via a point-wise temporal attention learner. Meanwhile, the processing of proximal accumulated residuals remains unchanged. This dual-scale learner design enables the model to simultaneously capture both temporal evolutionary patterns and the fine-grained high-frequency patterns that the base learner fails to capture. Furthermore, through a memory-enhanced cross-attention mechanism, residual representations are effectively refined, as noise and anomalies rarely maintain synchronous consistency between global trends and short-term fluctuations. Finally, beyond the reconstruction loss, we introduce a spatial consistency constraint based on KL divergence, that aligns the topological significance of each metric with its residual energy distribution to identify anomalies that deviate from the intrinsic systemic topology.

Our main contributions are summarized as follows:

- We propose RESTAD, a residual-enhanced spatio-temporal modeling method that redefines the role of residuals in multivariate KPI anomaly detection. RESTAD captures long-term trends from historical observations while repurposing the filtered residuals to extract local fine-grained details, which serve as a crucial supplement for robust reconstruction.
- We design a memory-enhanced cross-attention mechanism to maintain the synchronicity of short-term fluctuations with long-term trends.
- We present a spatial consistency contrastive strategy to amplify the discriminative gap between normal and anomalous patterns.
- Experimental results illustrate that RESTAD outperforms established baseline approaches in detection performance on four public real-world datasets.

II. Related Work

Time series anomaly detection has extensive applications in real-world AIOps. Early unsupervised approaches relied on clustering algorithms (e.g., ROCKA [6]) or recurrent neural networks (e.g., OmniAnomaly [7], InterFusion [5]) to capture basic temporal and inter-metric dependencies.

Recently, Transformer architectures [8] have demonstrated remarkable effectiveness in sequential data modeling across diverse domains, including natural language understanding [9, 10], audio signal processing, and computer vision tasks [11]. Leveraging the self-attention mechanism's capability to capture element-wise correlations adaptively, Transformers have been increasingly adopted for time series analysis to identify robust long-range temporal patterns [12–15]. In the context of time series anomaly detection, several Transformer-based approaches have emerged. PAFormer [16] presents an unsupervised parallel-attention framework integrating both global enhanced representation and local perception modules. Anomaly Transformer [17] amplifies the difference between global series-association and prior-association specific to anomalies through a renovated Anomaly-Attention mechanism based on association discrepancy, which guides the model to learn normal representations that are distinguishable from anomalies. Subsequently, several improvements [18–20] have been proposed. Beyond Transformers, emerging technologies such as diffusion models [21] and graph neural networks [22] have also been explored for addressing time series anomaly detection challenges.

III. Methodology

In this section, we detail the overall framework and technicalities of RESTAD, which comprises four modules: global spatial-attention base learner module, dual-scale temporal-attention learner module, spatio-temporal contrastive training module and anomaly detection module.

A. Overview

As shown in Fig. 1, we present the framework of RESTAD, which involves a bidirectional process of temporal decomposition and reconstruction. For a fixed-length window of multivariate KPIs $\{X_t, X_{t+1}, ..., X_{t+w-1}\}$ where $X_t \in \mathbb{R}^d$ is a d-dimensional vector, we first decompose it into predictable components and an unpredictable residual term. Specifically, predictable components are extracted from the previous observation window using a global spatial-attention base learner, while the residual term is directly derived by subtracting these predictable components from the input window. Logically, the current observation window can evolve from an initial state via an iterative residual-accumulation mechanism. The decomposition is expressed as follows:

$$X_{t:t+w-1} = f_{\text{enc}}(X_{t-1:t+w-2}) + \epsilon_{t:t+w-1} \tag{1}$$

where $\epsilon_{t:t+w-1}$ denotes the point-wise residual term, and $f_{\text{enc}}(\cdot)$ is a global spatial-attention network that captures explicit spatial correlations from global temporal features.

Reversely, through a recursive backtracking process, the previous observation window (with lag 1) can be reconstructed from a remote historical window (with lag τ) and the accumulated residuals. Here, we clarify a core property of residuals: although the residual at a single timestamp may appear as unpredictable noise, the accumulated historical residuals can retain the short-term fluctuation patterns of the predictable components. We reformulate (1), the predictable components can be further decomposed into short-term and long-term predictable components. Specifically, the short-term predictable components are learned via a local temporal-attention learner from the accumulated residuals, while the long-term predictable components are learned via a global temporal-attention learner from the remote historical window (with lag τ). The reconstruction process is expressed as follows:

$$\begin{aligned}\hat{X}_{t:t+w-1} &= f_{\text{dec}}(X_{t-\tau:t+w-1-\tau}) \\ &\quad + \sum_{k=1}^{\tau-1} \tilde{f}_{\text{dec}}(\epsilon_{t-k:t+w-1-k})\end{aligned} \tag{2}$$

where $f_{\text{dec}}(\cdot)$ denotes a global temporal-attention network designed to capture global temporal correlations, and $\tilde{f}_{\text{dec}}(\cdot)$ is a local temporal-attention network designed to capture local temporal correlations.

To suppress the interference of stochastic noise within the accumulated residuals and preserve only deterministic fluctuations, we augment $\tilde{f}_{\text{dec}}(\cdot)$ by incorporating a memory-enhanced cross-attention network. Inspired by the design of memory networks [23] , we use the long-term temporal

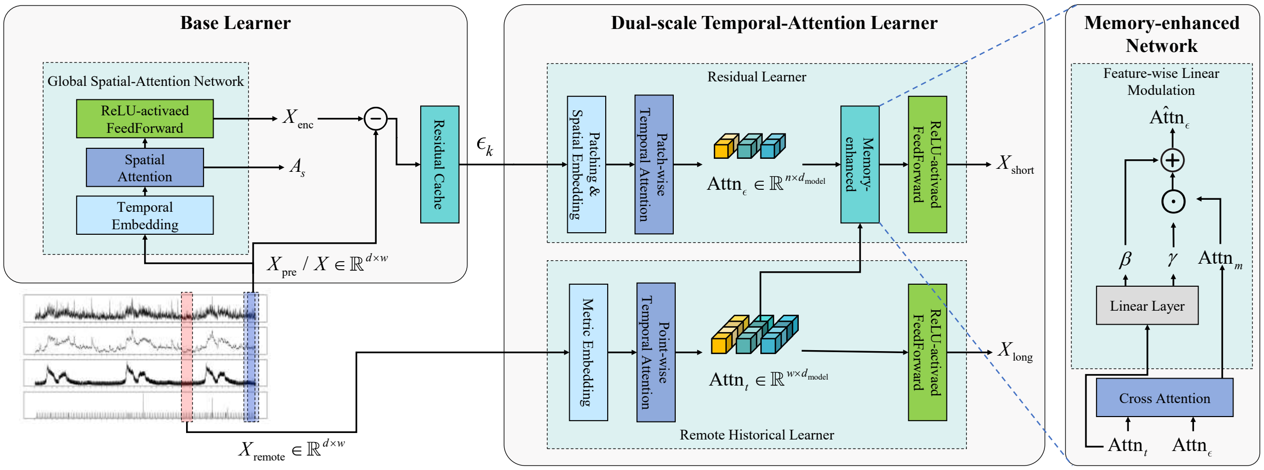


Fig. 1. RESTAD Framework

representations produced by $\phi_{\text{dec}}(\cdot)$ as the query to retrieve interpretable components from the short-term residual representations stored in the memory bank. This mechanism ensures that $\tilde{f}_{\text{dec}}(\cdot)$ extracts informative fluctuations that exhibit strong correlations with long-term temporal contexts, while unexplainable stochasticity is confined to the instantaneous residual window $\epsilon_{t:t+w-1}$. Accordingly, (2) is reformulated as follows:

$$\begin{aligned} H_{\text{long}} &= \phi_{\text{dec}}(X_{t-\tau:t+w-1-\tau}) \\ H_{\text{short}} &= \sum_{k=1}^{\tau-1} \tilde{\phi}_{\text{dec}}(\epsilon_{t-k:t+w-1-k}) \\ \hat{X}_{t:t+w-1} &= g_{\text{dec}}(H_{\text{long}}) + f_{\text{m}}(H_{\text{long}}, H_{\text{short}}) \end{aligned} \tag{3}$$

where H_{long} denotes long-term intermediate latent representations produced by the long-term feature extractor ϕ_{dec}, such that $f_{\text{dec}}(\cdot) = g_{\text{dec}} \circ \phi_{\text{dec}}$. H_{short} denotes short-term residual representations. $f_{\text{m}}(\cdot)$ denotes a memory-enhanced cross-attention network.

B. Global Spatial-Attention Base Learner

In this module, we first perform an embedding operation in the time dimension of the current observation window $X \in \mathbb{R}^{d \times w}$ and the previous observation windows $X_{\text{pre}} \in \mathbb{R}^{d \times w}$ to extract global temporal features. After obtaining the embeddings $X_{\text{emb}} \in \mathbb{R}^{d \times w_{\text{model}}}$ and $X_{\text{pemb}} \in \mathbb{R}^{d \times w_{\text{model}}}$, we adopt the attention mechanism from iTransformer [24] to explicitly model the spatial correlations that contain predictable structural information. The spatial-attention is formulated as follows:

$$\begin{aligned} Q_s = X_{\text{emb}} W_{Q_s},\ K_s = X_{\text{pemb}} W_{K_s},\ V_s = X_{\text{pemb}} W_{V_s} \\ A_s = \text{Softmax}\left(\frac{Q_s K_s^T}{\sqrt{w_{\text{model}}}}\right),\ \text{Attn}_s = A_s V_s \\ X_{\text{enc}} = \text{ReLU}(W_{e_1} \text{Attn}_s + b_{e_1}) W_{e_2} + b_{e_2} \end{aligned} \tag{4}$$

where $W_{Q_s}, W_{K_s}, W_{V_s} \in \mathbb{R}^{w_{\text{model}} \times w_{\text{model}}}$ are the transformation matrices for the query, key and value vector of spatial-attention, respectively. W_{e_1}, W_{e_2}, b_{e_1}, and b_{e_2} denote the transformation matrices and bias vectors of a ReLU-activated Feedforward Neural Network (FNN), which maps the time dimension back to its original window size. A_s represents the inter-metric topological graph of multivariate KPIs. X_{enc} represents the predictable components modeled by the spatial-attention base learner.

Then, we obtain the instantaneous residual window $\epsilon \in \mathbb{R}^{d \times w}$ via the subtraction model, these residuals are sequentially stored in a residual cache to implement the iterative accumulation mechanism, which is defined as follows:

C. Dual-scale Temporal-Attention Learner

1) Remote Historical Learner: In this module, we first perform an embedding operation in the metric dimension of the remote historical observation window $X_\tau \in \mathbb{R}^{w \times d}$ to obtain the embedding output $X_{\tau\text{emb}} \in \mathbb{R}^{w \times d_{\text{model}}}$. We adopt the attention mechanism from Transformer [8] to explicitly model the global temporal correlations that contain persistent trends and stable periodicity. The global temporal-attention is formulated as follows:

$$\begin{aligned} Q_t = X_{\tau\text{emb}} W_{Q_t},\ K_t = X_{\tau\text{emb}} W_{K_t},\ V_t = X_{\tau\text{emb}} W_{V_t} \\ \text{Attn}_t = \text{Softmax}\left(\frac{Q_t K_t^T}{\sqrt{d_{\text{model}}}}\right) V_t \\ X_{\text{long}} = \text{ReLU}(W_{d_1} \text{Attn}_t + b_{d_1}) W_{d_2} + b_{d_2} \end{aligned} \tag{5}$$

where W_{Q_t}, W_{K_t}, $W_{V_t} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ are the transformation matrices for the query, key and value vector of global temporal-attention, respectively. W_{d_1}, W_{d_2}, b_{d_1}, and b_{d_2} denote the transformation matrices and bias vectors of a ReLU-activated FFN, which maps the metric dimension back to its original numbers. X_{long} represents the long-term predictable components modeled by the global temporal-attention learner.

2) Residual Learner: This module comprises a local temporal-attention network and a memory-enhanced cross-attention network.

Local Temporal-Attention Network. In this network, residuals are treated as inputs. Instead of capturing point-wise correlations, we first partition the historical residual windows $\epsilon^k \in \mathbb{R}^{w\times d}$ into n non-overlapping patches. Each patch $\epsilon_i^k \in \mathbb{R}^{p\times d}$ encompasses p consecutive time steps, satisfying $n\times p = w$. Then, the in-patch and metric dimensions (p and d) are flattened into a unified vector space and we perform an embedding operation in the flattened dimension of each window to obtain the embedding outputs $\epsilon_{\text{emb}}^k \in \mathbb{R}^{n\times d_{\text{model}}}$. A key motivation for patching is that high-frequency oscillations often render individual residuals highly volatile and statistically unreliable. We similarly adopt the attention mechanism from Transformer [8] to explicitly model the local temporal correlations that encapsulate short-term fluctuation information. Note that all patches share the same parameters within the local temporal-attention network. Similarly, the local temporal-attention is formulated as follows:

$$\begin{aligned}
&Q_\epsilon^k = \epsilon_{\text{emb}}^k W_{Q_\epsilon},\ K_\epsilon^k = \epsilon_{\text{emb}}^k W_{K_\epsilon},\ V_\epsilon^k = \epsilon_{\text{emb}}^k W_{V_\epsilon} \\
&\text{Attn}_\epsilon^k = \text{Softmax}(\frac{Q_\epsilon^k (K_\epsilon^k)^T}{\sqrt{d_{\text{model}}}})V_\epsilon^k,\ \text{Attn}_\epsilon = \sum_{k=1}^{\tau-1}\text{Attn}_\epsilon^k
\end{aligned} \tag{6}$$

where $W_{Q_\epsilon}, W_{K_\epsilon}, W_{V_\epsilon} \in \mathbb{R}^{d_{\text{model}}\times d_{\text{model}}}$ are the transformation matrices for the query, key and value vector of local temporal-attention, respectively.

Memory-Enhanced Cross-Attention Network. After obtaining the patch-wise residual representations, we incorporate a memory-enhanced cross-attention mechanism to further distill predictable components from the residuals that exhibit synchronicity with the global temporal trends, thereby bolstering noise robustness. Specifically, the global temporal representations $\text{Attn}_t \in \mathbb{R}^{w\times d_{\text{model}}}$ serve as queries (Q_m), while the accumulated local residual representations $\text{Attn}_\epsilon \in \mathbb{R}^{n\times d_{\text{model}}}$ constitute the memory bank, which functions as keys (K_m) and values (V_m). The cross-attention is formulated as follows:

$$\begin{aligned}
&Q_m = \text{Attn}_t W_{Q_m}, K_m = \text{Attn}_\epsilon W_{K_m}, V_m = \text{Attn}_\epsilon W_{V_m} \\
&\text{Attn}_m = \text{Softmax}(\frac{Q_m K_m^T}{\sqrt{d_{\text{model}}}})V_m
\end{aligned} \tag{7}$$

where $W_{Q_m}, W_{K_m}, W_{V_m} \in \mathbb{R}^{d_{\text{model}}\times d_{\text{model}}}$ are the transformation matrices for the query, key and value vector of cross-attention, respectively.

Additionally, a Feature-wise Linear Modulation (FiLM) layer is integrated to modulate the residual output. By adaptively calibrating its magnitude and shifting its distribution, the FiLM layer ensures that the residual representations are effectively enhanced. The FiLM process is formulated as follows:

$$\begin{aligned}
\gamma, \beta &= \sigma(\text{Attn}_t) \\
\hat{\text{Attn}}_\epsilon &= \gamma \odot \text{Attn}_m + \beta
\end{aligned} \tag{8}$$

where γ, β are the scaling and shifting parameters derived from the global temporal representations Attn_t, and $\hat{\text{Attn}}_\epsilon$ denotes the calibrated residual output. Hence, we obtain the short-term predictable components X_{short} as modeled by the dual-scale temporal-attention learner, which is formulated as follows:

$$X_{\text{short}} = \text{ReLU}(W_{d_3}\hat{\text{Attn}}_\epsilon + b_{d_3})W_{d_4} + b_{d_4} \tag{9}$$

where W_{d_3}, W_{d_4}, b_{d_3}, and b_{d_4} denote the transformation matrices and bias vectors of a ReLU-activated FNN, which maps the latent space back to its original metric space.

Finally, according to (3), we obtain the reconstructed outputs as follows:

$$\hat{X} = X_{\text{long}} + X_{\text{short}} \tag{10}$$

D. Contrastive-Augmented Reconstruction Training Paradigm

In order to control the residuals under a structural prior of the spatial manifold, we introduce a spatial consistency constraint alongside the reconstruction loss. A straightforward insight is that in normal states, the inherent stability of the systemic topology forces stochastic fluctuations to propagate along fixed latent paths, but anomalous events can disrupt this stability. In essence, we regularize the accumulated residual energy distribution (P_ϵ) across the variable space to be similar to a structural prior (P_s) derived from the systemic topology. Specifically, we adopt entropy-weighted degree centrality as a measurement for this topological prior, and employ Kullback-Leibler (KL) divergence to quantify the discrepancy between the two distribution, which is formulated as follows:

$$\begin{aligned}
E^{(i)} &= \left|\bar{X}_{\text{short}}^{(i)}\right| \cdot \left(\sum_{j=1}^{d} A_s^{(i,j)} \cdot \left|\bar{X}_{\text{short}}^{(j)}\right|\right) \\
C^{(i)} &= \sigma\left(\sum_{j=1}^{d} A_s^{(i,j)}\right) \cdot e^{-H(A_s^{(i,\cdot)})} \\
P_\epsilon^{(i)} &= \frac{E^{(i)}}{\sum_{k=1}^{d} E^{(k)}},\ P_s^{(i)} = \frac{C^{(i)}}{\sum_{k=1}^{d} C^{(k)}} \\
L_{\text{contra}} &= \text{KL}(P_\epsilon \parallel P_{\text{s}}) + \text{KL}(P_s \parallel P_\epsilon)
\end{aligned} \tag{11}$$

where $E^{(i)}$ denotes the residual energy of dimension i. $C^{(i)}$ denotes the entropy-weighted degree centrality of dimension i. $H(A_s^{(i,\cdot)})$ is the Shannon entropy, defined as $H(A_s^{(i,\cdot)}) = -\sum_j A_s^{(i,j)} \log A_s^{(i,j)}$.

The total loss function of our model, which integrates contrastive constraints into the primary reconstruction criterion, is defined as follows:

$$L_{\text{total}} = \lambda_1 L_{\text{rec}} + \lambda_2 L_{\text{contra}} \tag{12}$$

where the reconstruction loss $L_{\text{rec}} = \left\|\hat{X} - X\right\|_{\text{F}}^2$ is equivalent to the mean squared error (MSE) under the Frobenius norm. λ_1 and λ_2 are the hyperparameter that balance the contributions of the two loss terms.

TABLE I
DATASET STATISTICS.

Dataset	Training	Testing	Dimension	Anomaly (%)
SMD	708405	708420	38	4.2
PSM	132481	87841	25	27.8
MSL	58317	73729	55	10.5
SMAP	135183	427617	25	12.8

TABLE II
ANOMALY EVENTS STATISTICS.

Dataset	Numbers	Duration (min)				
		Min	Max	Median	Average	Variance
SMD	327	2	3161	11	90.04	238.76
PSM	72	1	8861	5	338.63	1235.49
MSL	36	11	1141	121	215.72	275.26
SMAP	67	31	4218	131	816.36	1214.21

E. Anomaly Detection

During testing, we integrate a normalized KL similarity weight to amplify the anomaly scores for suspicious point-wise observations. Distinct from the spatial consistency constraint utilized during training, the KL divergence during testing is designed to assess temporal consistency by quantifying the distribution discrepancy between two temporal scales (X_{short} and X_{long}). A higher score indicates the input is poorly reconstructed by our model, reflecting significant temporal inconsistency. The final point-wise anomaly score is calculated as follows:

$$A_t = \text{Softmax}\left(\text{KL}(X_{\text{short}}|_t \| X_{\text{long}}|_t) + \text{KL}(X_{\text{long}}|_t \| X_{\text{short}}|_t)\right) \odot \left\| X_t - \hat{X}_t \right\|_{\text{F}}^2 \tag{13}$$

Specifically, if $A_t > \delta$, the time t is identified as an anomaly; otherwise (i.e., $A_t < \delta$), it is a normal point, where δ denotes a hyperparameter threshold, which is automatically determined according to the Peaks-Over-Threshold (POT) method [7].

IV. EXPERIMENTS

A. Datasets

We evaluate RESTAD on four real-world benchmark datasets: Server Machine Dataset (SMD) [7], Pooled Server Metrics (PSM) [25], Mars Science Laboratory (MSL) and Soil Moisture Active Passive (SMAP) [26]. As summarized in Table I and Table II, these datasets encompass diverse anomaly patterns. Specifically, SMD primarily consists of scattered point anomalies, whereas MSL and SMAP are characterized by sparse yet collective anomaly segments. In contrast, PSM encompasses both point anomalies and the longest-duration anomalies observed across the four datasets. This results in significantly higher uncertainty in its anomaly characteristics (as reflected by the variance). Consequently, our choice of benchmark datasets ensures a comprehensive coverage of diverse anomaly patterns.

B. Implementation Details

We implement the modeling, training, and evaluating of RESTAD in PyTorch with a single NVIDIA GeForce RTX 4090 24GB GPU. For simplicity and consistency, we standardize the hidden dimensions (d_{model} and w_{model}) of the embedding layers in all attention modules to 512, with the number of attention heads fixed at 4, the patch size for the residual learner is set to 10, and the historical lag τ for remote observation window is set to 3. The anomaly threshold ratio δ is configured as 0.5% for the SMD dataset and 1% for all other benchmarks. Finally, we use Adam optimizer and adopt an early stopping strategy. The initial learning rate is set to 10^{-4}, the batch size is set to 32, and the maximum number of iterations is 10 epochs.

C. Evaluation Metrics

To assess the effectiveness for anomaly detection, we employ the standard confusion matrix-based approach to compute three metrics: precision (P), recall (R), and F1-score (F1). Relevant formulas are listed as follows:

$$\text{P} = \frac{TP}{TP+FP}, \ \text{R} = \frac{TP}{TP+FN}, \ \text{F1} = \frac{2\times\text{P}\times\text{R}}{\text{P}+\text{R}} \tag{14}$$

where TP denotes true positive, TN denotes true negative, FP denotes false positive, and FN denotes false negative. Herein, positive instances refer to anomalous points, while negative instances refer to normal points. Note that we adopt the widely used point adjustment method [27] to rectify the initial TPs and FNs, thereby updating the aforementioned evaluation metrics.

D. Main Results

To verify the superior performance of our RESTAD, we compare it against nine state-of-the-art baselines, with all evaluations grounded in point-adjusted metrics. As observed from Table III, nine baselines include InterFusion [5], OmniAnomaly [7], MTAD-GAT [20], TranAD [28], Anomaly Transformer [17], MAN-QSM [18], DCDetector [19], DiG [22], and DConAD [29].

Based on the experimental results in Table III., our RESTAD achieves the best average F1-score of 95.91% across four benchmark datasets, outperforming nine baselines. Specifically, RESTAD demonstrates consistent advantages on the SMD, PSM, and SMAP datasets. Despite these datasets already being highly optimized by competitive baselines, RESTAD still yields incremental yet critical F1-score improvements of 1.55%, 0.45%, and 0.04%, respectively. The most significant improvement is observed on the SMD dataset, which is characterized by intricate non-Gaussian noise and high-frequency fluctuations. This leap primarily stems from our residual enhancement mechanism, which compensates for the "over-smoothing" effect commonly present in traditional methods. While existing Transformer-based models, such as Anomaly Transformer[17], DCDetector[19], TranAD[28], and DconAD[29], focus on capturing complex spatio-temporal dependencies, they often overlook the deterministic information in high-frequency residuals, resulting in high-frequency

TABLE III
PERFORMANCE COMPARISON.

Methods	SMD			PSM			SMAP			MSL		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
OmniAnomaly[7]	83.68	86.82	85.22	88.39	74.46	80.83	92.49	81.99	86.92	89.02	86.37	87.67
InterFusion[5]	87.02	85.43	86.22	83.61	83.45	83.52	89.77	88.52	89.14	81.28	92.70	86.62
MTAD-GAT[20]	76.67	40.72	53.19	98.14	58.59	73.37	94.40	55.92	70.23	91.75	84.07	87.74
TranAD[28]	90.60	72.73	80.69	1.00	58.56	73.87	91.17	70.90	79.77	78.78	34.33	47.82
Anomaly Transformer[17]	89.44	95.54	91.92	97.38	98.60	97.99	93.59	99.32	96.37	91.86	93.92	92.88
MAN-QSM[18]	89.07	94.21	91.57	97.42	98.49	97.95	93.62	99.54	96.49	92.06	97.93	94.90
DCDetector[19]	86.10	85.56	85.83	97.15	97.38	97.27	94.45	97.76	96.07	92.37	92.00	92.18
DiG[22]	90.10	93.74	91.88	95.16	97.58	96.36	94.26	93.81	94.03	92.09	95.15	93.59
DConAD[29]	80.87	91.94	87.05	96.00	98.17	97.07	94.50	98.79	96.60	91.68	99.30	**95.34**
RESTAD	89.50	97.80	**93.47**	97.47	99.43	**98.44**	93.65	99.83	**96.64**	92.19	98.13	95.07

TABLE IV
ABLATION STUDY RESULTS.

Model	SMD			PSM		
	P	R	F1	P	R	F1
w/o memory	89.17	94.83	91.91	96.83	97.14	96.98
w/o residual	88.43	93.18	90.74	96.11	95.37	95.74
w/o remote hist.	89.07	93.84	91.39	97.04	96.48	96.76
w/ temporal base	87.63	92.12	89.82	96.37	94.78	95.57
only recon loss	88.92	94.18	91.47	96.22	93.93	95.06
RESTAD	89.50	97.80	**93.47**	97.47	99.43	**98.44**

Note: w/o = without. 'remote hist.' = remote historical module. 'temporal base' = replacing spatial attention with temporal.

fluctuations being "masked" by dominant periodic trends. RESTAD effectively reconciles these time-frequency scales, ensuring that fine-grained details are actively recovered rather than discarded as noise. Furthermore, compared to frequency decomposition-based methods like [23], the residual mechanism preserves the original shape of those high-frequency fluctuations not explained by the base learner, thereby avoiding disruption of their time-frequency integrity.

E. Ablation Study

To verify the effectiveness and necessity of RESTAD as a whole, we conduct systematic ablation studies from two perspectives: the model architecture and the loss function. Considering computational efficiency and data representativeness, we perform these ablation experiments specifically on two primary benchmark datasets: SMD and PSM. The results are presented in Table IV.

Regarding the model architecture, we systematically exclude key components to assess their respective impacts on performance: the memory module within the residual learner (**w/o memory**), the entire residual learner itself (**w/o residual**), and the remote historical learner (**w/o remote hist.**). To further evaluate the specific contribution of spatial dependencies, we substitute the spatial attention network in the base learner with a temporal attention network (**w/ temporal base**). For the loss function, we design a variant that removes the KL divergence term (L_{contra}) and retains only the reconstruction loss (L_{rec}) under the Frobenius norm (**only recon loss**), thereby verifying the efficacy of the spatial consistency constraint. The observed performance degradation across all replacement and removal modules validates the necessity and efficacy of each constituent module.

F. Sensitivity Analysis

We conduct sensitivity experiments on four key hyperparameters: hidden dimensions (where we maintain $d_{\text{model}} = w_{\text{model}}$ for simplicity), input window size (w), patch size (p), and historical lag (τ), as shown in Fig. 2.

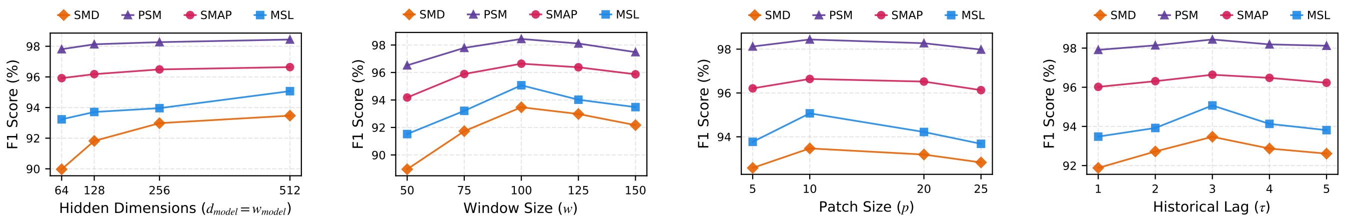
Generally, the hidden dimensions have a relatively minor impact on performance, indicating RESTAD's architectural efficiency. We set $d_{\text{model}} = w_{\text{model}} = 512$ to balance performance and complexity. In contrast, structural parameters significantly affect outcomes. A proper window size ($w = 100$) ensures sufficient perception of long-term trends, while an appropriate patch size ($p = 10$) maintains sensitivity to fine-grained residuals. Furthermore, the historical lag determines the temporal span for trend modeling alongside the scale of accumulated residuals. We observe that a relatively small lag ($\tau = 3$) achieves optimal performance, as deterministic high-frequency fluctuations typically manifest as transient impulses.

CONCLUSION

In this work, we propose a residual-enhanced spatio-temporal model that fully captures prominent spatial structure and stable trends while specifically extracting short-term fluctuations from accumulated residuals, thus enabling robust and accurate anomaly detection even amid complex noise. We introduce spatio-temporal attention mechanisms to refine the extraction of inherent patterns in multivariate KPIs. Furthermore, we incorporate a contrastive discrepancy loss to maximize the discriminability between anomalies and normal points. Comprehensive experiments on real-world datasets validate that our RESTAD delivers significant application value. In future work, we will consider more prior knowledge about anomalies to guide our model design, rather than just modeling general normal behaviors.

REFERENCES

[1] R. Mekuria, M. J. McGrath, V. Riccobene, V. Bayon-Molino, C. Tselios, J. Thomson, and A. Dobrodub, "Automated profiling of virtualized media processing functions using telemetry and machine learning," in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, pp. 150–161.



(a) Hidden Dimensions (b) Window Size (c) Patch Size (d) Historical Lag

Fig. 2. Sensitivity analysis of RESTAD with respect to four key hyperparameters: (a) hidden dimensions ($d_{model} = w_{model}$), (b) input window size (w), (c) patch size (p), and (d) historical lag (τ). The results are evaluated on SMD, PSM, SMAP, and MSL datasets using the F1 score.

[2] D. Liu, Y. Zhao, H. Xu, Y. Sun, D. Pei, J. Luo, X. Jing, and M. Feng, "Opprentice: Towards practical and automatic anomaly detection through machine learning," in *Proceedings of the 2015 internet measurement conference*, 2015, pp. 211–224.

[3] N. Zhao, J. Zhu, R. Liu, D. Liu, M. Zhang, and D. Pei, "Label-less: A semi-automatic labelling tool for kpi anomalies," in *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*. IEEE, 2019, pp. 1882–1890.

[4] O. Ibidunmoye, A.-R. Rezaie, and E. Elmroth, "Adaptive anomaly detection in performance metric streams," *IEEE Transactions on Network and Service Management*, vol. 15, no. 1, pp. 217–231, 2017.

[5] Z. Li, Y. Zhao, J. Han, Y. Su, R. Jiao, X. Wen, and D. Pei, "Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 3220–3230.

[6] Z. Li, Y. Zhao, R. Liu, and D. Pei, "Robust and rapid clustering of kpis for large-scale anomaly detection," in *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. IEEE, 2018, pp. 1–10.

[7] Y. Su, Y. Zhao, C. Niu, R. Liu, W. Sun, and D. Pei, "Robust anomaly detection for multivariate time series through stochastic recurrent neural network," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2828–2837.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[11] A. Wang, H. Chen, Z. Lin, J. Han, and G. Ding, "Repvit: Revisiting mobile cnn from vit perspective," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 15 909–15 920.

[12] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=rkgNKkHtvB

[13] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," *Advances in neural information processing systems*, vol. 32, 2019.

[14] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[15] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," *Advances in neural information processing systems*, vol. 34, pp. 22 419–22 430, 2021.

[16] N. Bai, X. Wang, R. Han, Q. Wang, and Z. Liu, "Paformer: Anomaly detection of time series with parallel-attention transformer," *IEEE transactions on neural networks and learning systems*, vol. 36, no. 2, pp. 3315–3328, 2023.

[17] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," in *ICLR*, 2022. [Online]. Available: https://openreview.net/forum?id=LzQQ89U1qm_

[18] J. Zhong, E. Zuo, C. Chen, C. Chen, J. Yan, T. Li, and X. Lv, "A masked attention network with query sparsity measurement for time series anomaly detection," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 2741–2746.

[19] Y. Yang, C. Zhang, T. Zhou, Q. Wen, and L. Sun, "Dcdetector: Dual attention contrastive representation learning for time series anomaly detection," in *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, 2023, pp. 3033–3045.

[20] H. Zhao, Y. Wang, J. Duan, C. Huang, D. Cao, Y. Tong, B. Xu, J. Bai, J. Tong, and Q. Zhang, "Multivariate time-series anomaly detection via graph attention network," in *2020 IEEE international conference on data mining (ICDM)*. IEEE, 2020, pp. 841–850.

[21] C. Wang, Z. Zhuang, Q. Qi, J. Wang, X. Wang, H. Sun, and J. Liao, "Drift doesn't matter: Dynamic decomposition with diffusion reconstruction for unstable multivariate time series anomaly detection," *Advances in neural information processing systems*, vol. 36, pp. 10 758–10 774, 2023.

[22] D. Lan, G. Zhang, and R. Guo, "Diffusion graph model for time series anomaly detection via anomaly-aware graph sparsification and augmentation," in *Companion Proceedings of the ACM on Web Conference 2025*, 2025, pp. 2207–2214.

[23] J. Song, K. Kim, J. Oh, and S. Cho, "MEMTO: Memory-guided transformer for multivariate time series anomaly detection," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=UFW67uduJd

[24] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "itransformer: Inverted transformers are effective for time series forecasting," in *ICLR*, 2024. [Online]. Available: https://openreview.net/forum?id=JePfAI8fah

[25] A. Abdulaal, Z. Liu, and T. Lancewicki, "Practical approach to asynchronous multivariate time series anomaly detection and localization," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 2485–2494.

[26] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 387–395.

[27] H. Xu, W. Chen, N. Zhao, Z. Li, J. Bu, Z. Li, Y. Liu, Y. Zhao, D. Pei, Y. Feng *et al.*, "Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications," in *Proceedings of the 2018 world wide web conference*, 2018, pp. 187–196.

[28] S. Tuli, G. Casale, and N. R. Jennings, "TranAD: deep transformer networks for anomaly detection in multivariate time series data," *Proc. VLDB Endow.*, vol. 15, no. 6, pp. 1201–1214, Feb. 2022.

[29] W. Zhang, X. Lin, W. Yu, G. Yao, J. Zhong, Y. Li, R. Han, S. Xu, H. Shi, and C. Luo, "Dconad: A differencing-based contrastive representation learning framework for time series anomaly detection," in *2025 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2025, pp. 1–8.