A REPRESENTATION LEARNING PERSPECTIVE ON REVISITING RECONSTRUCTION-BASED TIME-SERIES ANOMALY DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstruction-based approaches are one of the mainstream paradigms for time-series anomaly detection. Existing studies typically improve anomaly detection performance by designing sophisticated network architectures or introducing prior biases to capture specific characteristics of time-series data. However, we find that these approaches still have limitations in effectively capturing data characteristics and often incur substantial computational costs. To address these issues, we introduce a Transformer-based patch-level anomaly detection model that focuses on learning global representations of time-series data. By combining point-wise and interval-wise anomaly scores, our method achieves anomaly detection performance comparable to or even better than that of mainstream approaches, without requiring complex network architectures or manually designed prior biases. It also achieves a favorable trade-off between detection performance and computational cost. Furthermore, we investigate the scalability and transferability, providing a new perspective on reconstruction-based time-series anomaly detection and suggesting that increasingly complex network architectures may not be necessary for effective anomaly detection.

1 INTRODUCTION

Time series data, characterized as a temporally ordered sequence of observations, widely exists across various domains such as industrial control, KPI (Zamanzadeh Darban et al., 2024; Wen et al., 2022; Dai et al., 2021; Huang et al., 2022). Within this context, time series anomaly detection (TSAD) stands as a critical task in time series analysis, aiming to identify observations or patterns that deviate significantly from established normal behaviors. Recently, deep learning-based approaches, especially reconstruction-based models, have been extensively developed to TSAD. Compared with alternative detection approaches, reconstruction-based paradigms are more straightforward and unified, offer higher computational efficiency, and demonstrate superior detection performance (Yahya et al., 2025). Reconstruction-based anomaly detection methods (Xu et al., 2021; Feng et al., 2024; Wu et al., 2025; Obata et al., 2025) aim to learn appropriate representations of normal samples in the latent space from training data, and then map them back to the data space through reconstruction. Anomalies are subsequently identified based on the discrepancy between the test data and their reconstructed counterparts. Previous studies have followed this line of research (Park et al., 2018; Audibert et al., 2020; Tuli et al., 2022; Su et al., 2019) to learn robust features from data. Recent studies have increasingly focused on incorporating tailored inductive biases and sophisticated architectural designs to capture more complex features in time series data. For instance, TimesNet (Wu et al., 2022) model intra-period and inter-period dependencies in the time domain to capture temporal evolution; TSINR (Li et al., 2025) extract continuity features by learning low-frequency temporal components; Catch (Wu et al., 2025) leverage frequency-domain information to characterize normal behavioral patterns within the data; SimAD (Zhong et al., 2025) proposes an enhanced attention mechanism that progressively captures the feature discrepancies between normal and anomalous samples in a layer-wise manner.

While these approaches provide multi-faceted understandings of "normality" from various angles, they typically introduce complex network architectures or prior biases to capture specific characteristics of the data, which potentially limits the model's generalization and robustness. For example, a

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

phenomenon is the model's tendency to overfit to anomalous samples during the reconstruction process (see fig. 1). This observation prompts us to reconsider whether existing reconstruction-based mechanisms succeed largely because anomalies are inherently more difficult to reconstruct at the level of low-level numerical semantics. In other words, we are curiosity with whether these "so-called diverse feature representations" can genuinely improve the performance of reconstruction-based anomaly detection. Compared with the complex textures, shadows, and other visual features present

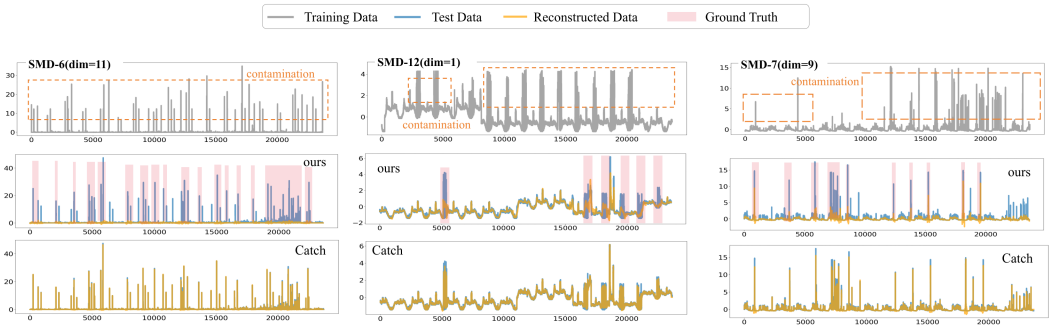


Figure 1: Comparison with Catch (Wu et al., 2025). Even when the training data are heavily contaminated by anomalies, our method captures the essential characteristics of normal patterns, whereas Catch tends to be influenced by anomalous observations.

in images, time-series data typically exhibit relatively simple patterns, primarily characterized by trends and seasonality (Wen et al., 2019; Zhou et al., 2022). Therefore, we argue that complex network architectures or inductive biases designed to capture particular patterns may not be the key factors for improving the performance of reconstruction-based anomaly detection, especially in terms of generalization and transferability. Instead, this raises a fundamental question: to what extent does representation learning contribute to the effectiveness of reconstruction-based time-series anomaly detection? Our work investigates this question by examining the role of representation learning in reconstruction-based time-series anomaly detection. Inspired by Vision Transformer (Dosovitskiy et al., 2020), we tokenize the time series into patches and map them into latent embeddings via linear projection. The patch embeddings, together with a learnable CLS token that serves as a global representation of the sequence, are processed by a Transformer encoder backbone. For anomaly score, beyond point-wise reconstruction errors, we additionally measure interval-level reconstruction discrepancies in the frequency domain (Nam et al., 2024), enabling the detection of anomalous patterns.

Experimental results demonstrate that (see fig. 2), without relying on sophisticated architectural designs or manually introduced inductive biases, our method achieves performance comparable to or surpassing existing state-of-the-art approaches, achieving a strong trade-off between detection performance and computational cost. Furthermore, we find that performance does not necessarily improve with increasing model capacity or training data scale. This observation suggests that, in reconstruction-based anomaly detection, simply scaling up models and datasets may not be sufficient to enhance anomaly detection performance. Regarding transferability, the removal of domain-specific inductive biases enables the learned representations to generalize more effectively across datasets. Notably, our method achieves competitive detection performance with only linear probing, even compared with models that are fully retrained on the target dataset.

Our contributions are three-fold:

- We identify and study the limitations of existing reconstruction-based time series anomaly detection methods, particularly their reliance on complex architectures and inductive biases.
- We find that a simple global representations combined with an appropriate anomaly scoring strategy can achieve effective anomaly detection.
- We revisit the role of representation learning in reconstruction-based anomaly detection and investigate its scalability and transferability across datasets.

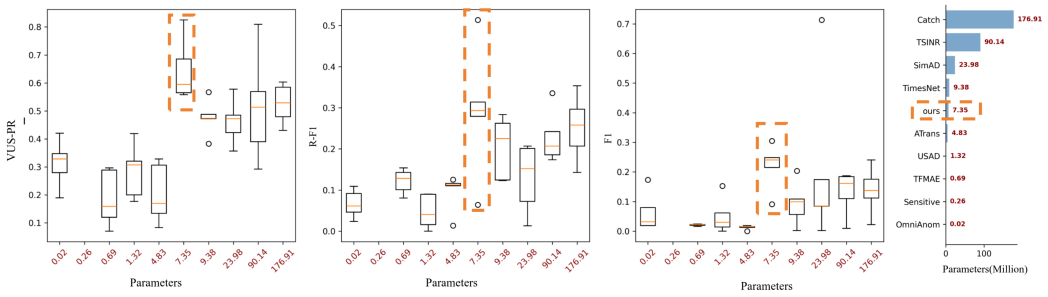


Figure 2: Comparison of performance metrics and model parameter counts for different methods across five datasets. The box plots illustrate the distribution of performance metrics achieved by each method on the five datasets, while the bar charts present the corresponding numbers of model parameters.

2 Related Work

2.1 Reconstruction based Time Series Anomaly Detection

The core of reconstruction-based methods lies in the compression representation and decoding reconstruction of time series, where anomaly scores are measured based on reconstruction errors. These methods facilitate feature learning from unlabeled data without relying on ground-truth labels and demonstrate significant advantages in capturing local outliers, such as contextual anomalies (Yahya et al., 2025). Reconstruction-based methods can be further classified into several architectural paradigms, including AE, VAE, and GAN-based models (Zamanzadeh Darban et al., 2024; Yahya et al., 2025), as exemplified by LSTM-VAE (Park et al., 2018), DAGMM (Zong et al., 2018), OmniAnomaly (Su et al., 2019), and TranAD (Tuli et al., 2022).

Recently, researchers have been dedicated to designing methods that characterize the key features of time-series data from multiple dimensions. For instance, TimesNet (Wu et al., 2022) captures continuous temporal characteristics by explicitly modeling both inter-period and intra-period variations. Observing that distinct anomaly types and cross-channel correlations manifest uniquely in the frequency domain, Catch (Wu et al., 2025) leverages frequency-domain information to learn underlying ”normality” hypotheses. SensitiveHUE (Feng et al., 2024) proposes a feature-removal strategy to eliminate statistical features from the data, while redesigning a Negative Log-Likelihood (NLL) loss to mitigate the adverse effects of noise on data modeling. TSINR (Li et al., 2025) capitalizes on the spectral bias inherent in Implicit Neural Representations (INR), which naturally prioritizes fitting low-frequency signals to preserve and represent data continuity. SimAD (Zhong et al., 2025) incorporates ”Embed Patch” tokens to serve as multi-level feature prototypes, effectively characterizing data representations across different granularities. While these methods provide a diverse understanding of ”normality” from different angles, certain limitations and biases remain when it comes to fully capturing global data characteristics (Wang et al., 2024).

2.2 Representation Learning

Representation learning aims to learn more robust, discriminative, and generalizable feature representations. It has achieved remarkable progress in both computer vision and natural language processing (Devlin et al., 2019; Dosovitskiy et al., 2020; He et al., 2022; Chen et al., 2020). Recently, the core principles of representation learning have been increasingly applied to time-series representation tasks. For example, PatchTST employs a masked autoencoding strategy to learn patch-based representations for time series prediction (Nie, 2022). TS2Vec adopts random cropping and timestamp masking strategies to perform contrastive learning across different augmented views (Yue et al., 2022). TS-TCC designs temporal augmentation strategies with varying degrees of perturbation to construct positive and negative sample pairs for contrastive learning (Eldele et al., 2021). TF-C is based on the assumption that the same sample exhibits higher similarity in both the time and frequency domains compared with other samples; accordingly, it designs augmentation strategies in both domains to construct positive and negative pairs for contrastive learning (Zhang et al., 2022). Leveraging the

acquired feature representations, existing approaches primarily verify their efficacy on downstream applications like time-series classification and forecasting. However, dedicated investigations into learning more robust and generalizable features tailored for time series anomaly detection remain under-explored.

3 METHOD

3.1 PROBLEM STATEMENT

We consider the following problem: given a multivariate time series $X \in \mathbb{R}^{M \times N}$ with M channels and N steps. We denote $X^m = \{X_1^m, ..., X_N^m\} \in \mathbb{R}^N$ as the univariate time series of the m-th channel, where $m = 1, ..., M$. Similarly, $X_i = \{X_i^1, ..., X_i^M\} \in \mathbb{R}^M$ as the multivariate data point at the i-th time step, where $i = 1, ..., N$. Our goal is to reconstruct X (denoted as $\hat{X}$) while preserving its normal temporal patterns, such as trends and seasonality. A anomaly score $f(X_i, \hat{X}_i)$ determine whether data point X_i is anomaly or not.

3.2 MODEL STRUCTURE

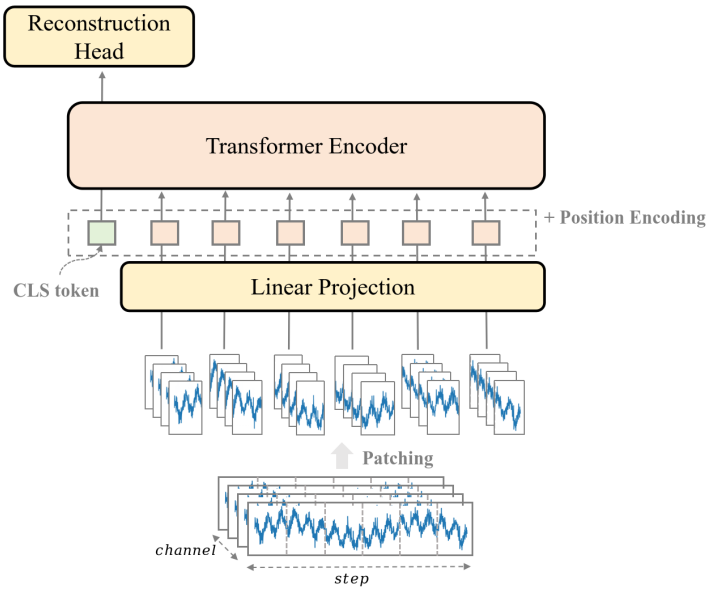


Figure 3: Model Overview. Inspired by ViT (Dosovitskiy et al., 2020), we embed non-overlapping patches and feed them into the Transformer encoder. A learnable CLS token is prepended to the patch embeddings to serve as a global representation of the sequence. A linear layer projects the CLS token back to the original data space.

An overview of the model is depicted in fig. 3. To mitigate the distribution shift between the training and testing data, we normalize X along the temporal dimension before patching and denormalize the model outputs afterward (Kim et al., 2021):

$$X^m = \gamma \frac{X^m - \mathbb{E}[X^m]}{\sqrt{Var[X^m] + \epsilon}} + \beta, \quad m = 1, ..., M \tag{1}$$

where γ and β are learnable affine parameters, and ϵ is a small constant to prevent division by zero. As a result, the normalized X can have a more consistent mean and variance, where the non-stationary information is reduced.

We Then divide X into a sequence of non-overlapping 2D patches $\{x_p^1, ..., x_p^L\} \in \mathbb{R}^{L \times (P \times M)}$, where P denotes the patch size and $L = \lceil \frac{N}{P} \rceil$. To ensure complete partitioning, we pad X along the time

dimension before patching. With these patches, the number of input tokens can be reduced, thereby decreasing memory consumption and computational complexity (Nie, 2022).

We then flatten the patches and project them into a D-dimensional latent space through a linear embedding layer, yielding a sequence of patch embeddings. Following ViT (Dosovitskiy et al., 2020) and BERT (Devlin et al., 2019), we prepend a learnable embedding token, denoted as x_{cls}, to the patch embeddings as the global representation of the time series. We expect this CLS token to aggregate global contextual information through self-attention and capture high-level semantic features of the data during the reconstruction process.

Then, We adopt a vanilla Transformer (Vaswani et al., 2017) encoder as the backbone to learn the feature representation $h_0, h_1, ..., h_L$, where $h_i \in \mathbb{R}^D$. A Linear Layer is applied to h_0 to obtain the final reconstruction $\hat{X} \in \mathbb{R}^{M \times N}$. We further observe that, despite the better expressive power of MLP prediction heads, a simple linear layer is sufficient for reconstruction and yields better efficiency and scalability. The overall architecture can be formulated as follows:

$$x_p^i = Patchify(X), i = 1, \ldots, L \tag{2}$$

$$z = [x_{cls}; x_p^1 W_e; ...; x_p^L W_e] + E_{pos}, \tag{3}$$

$$h_0, h_1, .., h_L = TransformerEncoder(z), \tag{4}$$

$$\hat{X} = LinearLayer(h_0) \tag{5}$$

where x_{cls} is a learnable token, $W_e \in \mathbb{R}^{(P \times M) \times D}$ is the linear projection matrix, and $E_{pos} \in \mathbb{R}^{(L+1) \times D}$ is the learnable positional embedding to capture the temporal order of the patches.

3.3 Anomaly Score

Point-wise reconstruction errors are commonly used as anomaly scores for anomaly detection. However, even when the model successfully captures normal patterns, point-wise reconstruction errors may still fail to identify certain types of collective anomalies (Yahya et al., 2025). To address this limitation, in addition to computing point-wise reconstruction errors, we further measure reconstruction errors in the frequency domain. Specifically, both the original and reconstructed signals are transformed into the frequency domain via the Fourier transform, and the discrepancy between their frequency representations is used as an additional anomaly score (Nam et al., 2024).

We observe that point-wise and frequency-domain reconstruction errors often exhibit substantially different numerical scales. Directly combining them may cause one component, typically the point-wise reconstruction error, to dominate the final anomaly score, thereby degrading detection performance. To alleviate this issue, we first apply a logarithmic transformation to normalize the scale of each score and then perform max pooling along the channel dimension to enhance discriminative differences before aggregation:

$$score_p(X_i) = \log(X_i - \hat{X}_i)^2 \tag{6}$$

$$score_f(X_i) = \frac{1}{N} \sum_{i=1}^{N} (|fft(X_i) - fft(\hat{X}_i)|) \tag{7}$$

$$score = score_p(X_i) + score_f(X_i) \tag{8}$$

4 Experiment

4.1 Setup

Baselines. We consider the following methods as baselines for comparison, including seven reconstruction-based anomaly detection methods: Catch (Wu et al., 2025), TSINR (Li et al., 2025), SimAD (Zhong et al., 2025), SensitiveHUE (Feng et al., 2024), TimesNet (Wu et al., 2022), USAD (Audibert et al., 2020), and OmniAnomaly (Su et al., 2019). To provide a more comprehensive evaluation, we additionally include two feature-based anomaly detection methods, AnomalyTransformer (Xu et al., 2021) and TFMAE (Fang et al., 2024), as well as two traditional machine learning methods LOF (Breunig et al., 2000) and HBOS (Goldstein & Dengel, 2012). All methods are implemented

based on their publicly available code repositories, and the hyperparameter settings are kept consistent with the original implementations.

Datasets. Following the dataset settings adopted by most previous studies, we evaluate the performance of our method on five real-world datasets, including SMD (Su et al., 2019), MSL (Su et al., 2019), SMAP (Su et al., 2019), SWaT (Mathur & Tippenhauer, 2016), and PSM (Abdulaal et al., 2021). These datasets cover time-series data from various domains, including server monitoring systems, satellite systems, and industrial control systems, with diverse anomaly types and distribution characteristics. During data processing, all baseline methods adopt the same preprocessing procedures, including data normalization and missing value handling, to ensure fairness and comparability of the experiments.

Implementation Details. By default, we employ a two-layer Transformer encoder with head number 4 and dimension of latent space 128. The feed-forward network consists of two linear layers with ReLU activations. Dropout with probability 0.1 is applied in the encoders. The input sliding window size follows the setting of Catch (Wu et al., 2025). For all experiments, we use an initial learning rate of 0.001, train the model for 5 epochs, and set the batch size to 128. We adopt the Adam optimizer (Kingma & Ba, 2014) with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Meanwhile, a StepLR scheduler is employed to adjust the learning rate during training. The detailed model configurations and evaluation hyperparameter settings are provided in the supplementary material. The code and datasets will be publicly available.

Metrics. We employ both range-wise and point-wise metrics to evaluate the performance of different methods. For the range-wise evaluation, we adopt VUS-PR (Paparrizos et al., 2022) and Range-F1 (Tatbul et al., 2018). Among them, VUS-PR has been demonstrated to be a robust (insensitive to temporal delays), accurate (unbiased and effective across different scenarios), and fair (consistent under similar conditions) evaluation metric (Liu & Paparrizos, 2024). For the point-wise evaluation, we use the F1 score. To ensure a fair comparison, we adopt the same evaluation codebase[1] and identical evaluation settings for all methods.

4.2 RESULTS

Table 1: Average A-R (AUC-ROC), R-A-R (R-AUC-ROC) and V-ROC (VUS-ROC) accuracy measures for all datasets. The best results are highlighted in bold, and the second-best results are underlined.

Method	SMD			PSM			SWaT			MSL			SMAP		
	VUS-PR	F1	R-F1	V-PR	F1	R-F1	VUS-PR	F1	R-F1	V-PR	F1	V-F1	V-PR	F1	V-F1
ours	**0.686**	**0.308**	**0.307**	0.600	**0.094**	**0.507**	**0.824**	0.280	0.064	**0.560**	**0.243**	**0.309**	**0.555**	**0.218**	**0.280**
Catch	0.584	0.175	0.142	**0.603**	0.022	0.353	0.479	0.112	**0.206**	0.528	0.240	0.296	0.430	0.137	0.258
SimAD	0.485	0.174	0.201	0.472	0.002	0.206	0.578	**0.713**	0.012	0.422	0.085	0.072	0.357	0.084	0.151
TSINR	0.513	0.161	0.206	0.569	0.009	0.335	0.809	0.183	0.185	0.390	0.186	0.242	0.292	0.109	0.173
Sensitive	0.469	0.119	0.092	0.540	0.092	0.081	0.480	0.100	0.170	0.451	0.170	0.257	0.367	0.108	0.202
TimesNet	0.567	0.099	0.225	0.472	0.002	0.124	0.382	0.056	0.122	0.487	0.203	0.283	0.473	0.108	0.262
Omni	0.189	0.173	0.091	0.420	0.031	0.046	0.279	0.019	0.061	0.347	0.018	0.023	0.328	0.079	0.109
USAD	0.176	0.153	0.090	0.419	0.030	0.040	0.200	0.000	0.000	0.306	0.014	0.016	0.320	0.061	0.089
TFMAE	0.070	0.023	0.080	0.288	0.022	0.154	0.296	0.019	0.142	0.159	0.024	0.128	0.120	0.015	0.100
ATrans	0.082	0.062	0.125	0.328	0.001	0.013	0.306	0.015	0.110	0.169	0.015	0.116	0.134	0.018	0.112
LOF	0.169	0.125	0.203	0.380	0.034	0.401	0.459	0.024	0.142	0.202	0.063	0.189	0.138	0.029	0.174
HBOS	0.277	0.217	0.151	0.490	0.045	0.141	0.760	0.148	0.133	0.245	0.108	0.140	0.217	0.055	0.127

Main results. Table 1 and fig. 2 summarize the results of our method and the baselines across all datasets. Overall, our method consistently outperforms the baselines in terms of both range-wise and point-wise metrics across most datasets, it deconstrate the effectiveness in detecting both point anomalies and pattern anomalies. On the SMD dataset, We observe the presence of anomaly contamination in the training data, which can potentially hinder the model's ability to learn normal patterns (see fig. 1). Due to the ability to capture global contextual information and learning robust representations for anomaly detection, our method performs well even in the presence of anomaly contamination within the long-range dataset. On the SWaT dataset, our method achieves competitive performance in terms of VUS-PR, while SimAD attains the highest F1 score. This discrepancy may stem from the fact that SimAD adopts a more complex architecture with a larger number of parameters, enabling it to model fine-grained local patterns more effectively. As a result, SimAD

[1] https://github.com/thedatumorg/VUS/tree/main

exhibits superior performance on point-wise metrics. In contrast, our method is designed to capture global contextual information and long-range temporal dependencies, which contributes to its stronger performance on range-based metrics.

Sensitiveity analysis. Figure 5 presents the sensitivity analysis results of our method with respect to the input window size and patch size. We evaluate window sizes of 128, 192, 256, and 320. The patch sizes are set to 1 (without patching operation), 16, 32, 48, and 64.

Specifically, all three evaluation metrics on SMD consistently decline with larger window sizes, whereas SMAP and MSL exhibit a non-monotonic trend. As illustrated in fig. 4, we further analyze the anomaly length distributions of SMAP, MSL, and SMD. The statistics reveal that the anomaly segments in SMD are generally shorter and more sporadic than those in SMAP and MSL. Therefore, for SMAP and MSL, performance degradation is observed only at window sizes with relatively low occurrence frequencies (e.g., 128 and 256). In contrast, for SMD, excessively large windows weaken the model's ability to focus on local details, leading to degraded performance in detecting short-range anomalies. In contrast, the model is less sensitive to variations in patch size, exhibiting more stable performance across different patch sizes.

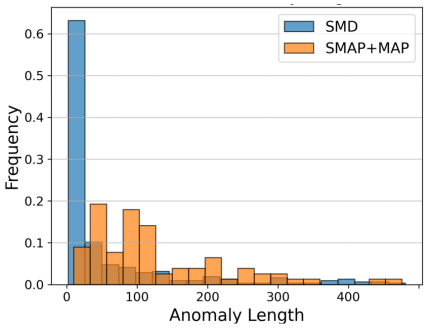


Figure 4: distribution of anomaly lengths in SMAP, MSL, and SMD datasets.

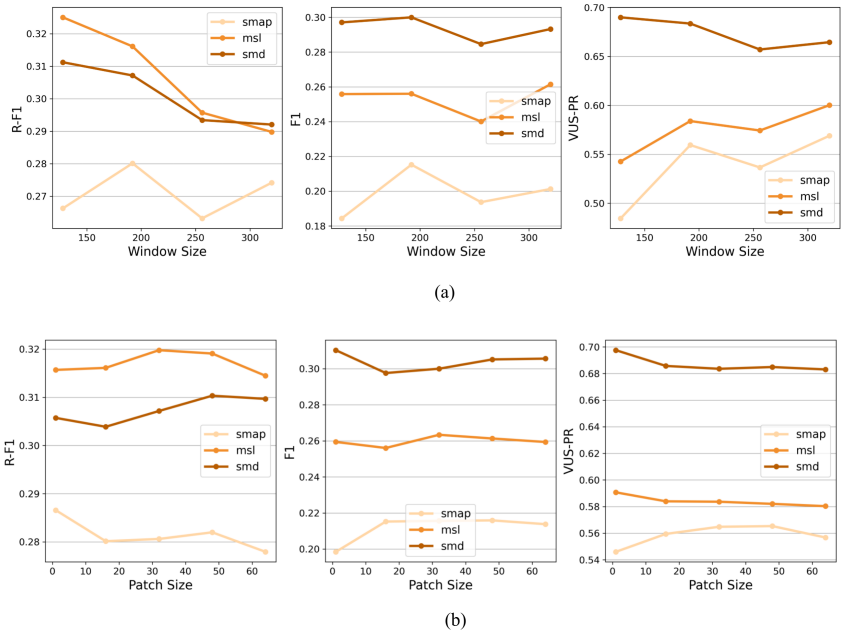


Figure 5: Sensitivity Analysis of Patch Size and Window Size.

Ablation. Table 2 presents the results of our ablation study, which evaluates the contribution of different components to the overall performance of our method. Specifically, we investigate: (1) removing the frequency-domain reconstruction error $score_f$ from the anomaly score, and (2) replacing the linear reconstruction head with an MLP. The results demonstrate that incorporating the frequency-domain reconstruction error consistently improves anomaly detection performance. Therefore, rather than pursuing increasingly complex representations, it is equally important to consider how to effectively construct anomaly scores from the learned representations. Moreover, compared

with an MLP, the linear head yields comparable or better results while requiring substantially fewer parameters, making it a more efficient design choice.

Table 2: Ablation study on five benchmark datasets.

Method	SMAP			PSM			SMD			SWaT			MSL		
	VUS-PR	R-F1	F1	VUS-PR	R-F1	F1	VUS-PR	R-F1	F1	VUS-PR	R-F1	F1	VUS-PR	R-F1	F1
w/o $score_f$	0.469	0.259	0.181	0.474	0.447	0.087	0.509	0.258	0.307	0.715	**0.115**	0.208	0.507	0.283	0.230
replace linear projection	**0.560**	**0.287**	**0.224**	0.589	0.497	0.088	0.685	0.305	0.299	0.818	0.065	**0.343**	0.559	**0.320**	0.251
ours	0.555	0.280	0.218	**0.600**	**0.507**	**0.094**	**0.686**	**0.307**	**0.308**	**0.825**	0.064	0.280	**0.560**	0.309	**0.243**

Table 3: Hyperparameter configurations for the three model scales.

Model	# Layers	# Heads	Hidden size	FFN size	Params
base	2	4	128	256	1.3M
medium	6	8	256	512	5.3M
large	12	16	512	1024	29.5M

Scalable. We further investigate the influence of training dataset scale and model size on performance, aiming to explore whether scaling laws hold for time-series anomaly detection tasks. We conduct experiments on the PSM and SMD datasets due to their relatively large data scales. Specifically, we evaluate the model using 0.25, 0.50, 0.75, and 1.0 proportions of the training data while keeping the test set unchanged. For the model size, as shown in table 3, we adopt three groups of parameter configurations to represent different model scales, where the *base* setting corresponds to the default configuration used in this paper.

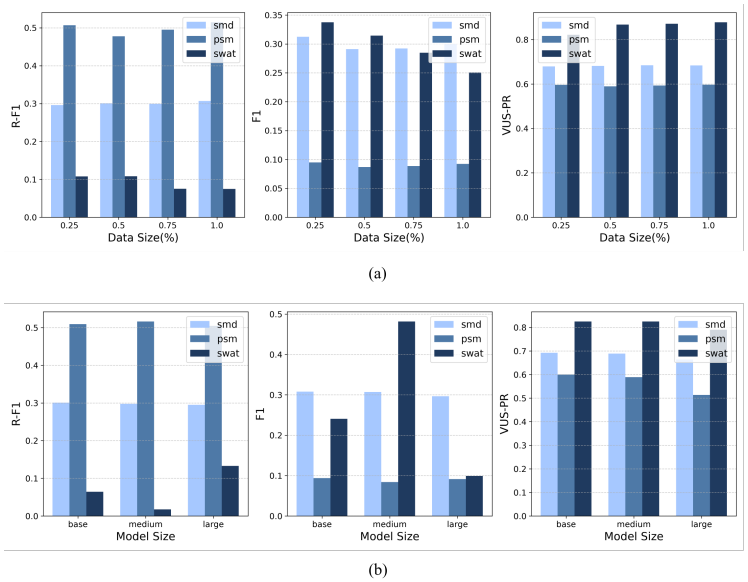


Figure 6: Performance of different model sizes and training data scales.

Figure 6 illustrates the relationship between metrics, dataset scale, and model size. It can be observed that the model does not exhibit the expected scaling law phenomenon, as the evaluation metrics remain nearly unchanged and even slightly decline in some cases as the model or dataset scale increases. This observation suggests that TSAD may not necessarily benefit from excessively large datasets or overly deep models. Unlike image, which contain highly complex patterns such as edges, textures, and shadows, time series are usually more periodic and easier to characterize. On the contrary, excessively large models may increase the risk of overfitting and consequently degrade performance. This also explains why, compared with simple machine learning or statistical methods,

deep learning approaches incur substantially higher computational costs on training data while still failing to demonstrate clear absolute competitive advantages in performance (Schmidl et al., 2022).

Transfer Learning. In real-world applications, retraining a model for every new domain is offten costly. Therefore, we further investigate whether a generalizable feature representation learned for anomaly detection can be effectively transferred across different datasets. Given the substantial variation in channel dimensionality across datasets, conducting unified training over multiple datasets is challenging. Therefore, we select four univariate time series to evaluate the transferability of the pretrained model: WSD, AIOps (Li et al., 2022), Yahoo, and NAB (Lavin & Ahmad, 2015). AIOps and WSD consist of key performance indicator (KPI) time series collected from large-scale Internet services, whereas Yahoo and NAB contain time series from a diverse range of application domains.

Regarding the training procedure, we first pre-train the model on the AIOps dataset, and then perform linear probe evaluation on the NAB, Yahoo, and WSD datasets. In addition, we select three competitive baselines with prominent performance for comparison: Catch (Wu et al., 2025), TSINR (Li et al., 2025), SimAD (Zhong et al., 2025) and TimesNet (Wu et al., 2022). These benchmarking methods are trained and evaluated directly on the NAB, Yahoo, and WSD datasets. It is worth noting that only our pre-trained model undergoes linear probe evaluation on the Yahoo, NAB, and WSD datasets, whereas all baseline methods are retrained and evaluated directly on these benchmarks. The experimental results are summarized in table 4. Aside from substantially mitigating the heavy computational overhead associated with full model retraining, our approach consistently achieves superior performance across all three datasets. This highlights the capacity of our model to learn generalizable feature representations for effective cross-domain anomaly detection.

Table 4: Transfer to Yahoo, NAB and WSD datasets. Times(s) is the training time of each method on the three datasets.

Method	Yahoo				NAB				WSD			
	VUS-PR	R-F1	F1	Times	VUS-PR	R-F1	F1	Times	VUS-PR	R-F1	F1	Times
Ours	**0.737**	**0.705**	**0.702**	4.24s	**0.311**	0.244	**0.164**	8.20s	**0.648**	0.311	**0.338**	8.40s
Catch	0.735	0.693	0.687	9.10s	0.290	0.237	0.057	57.79s	0.456	**0.370**	0.303	71.66s
TSINR	0.404	0.213	0.212	57.14s	0.148	0.251	0.049	54.95s	0.300	0.279	0.239	62.04s
SimAD	0.408	0.237	0.239	9.38s	0.194	**0.262**	0.071	30.36s	0.208	0.174	0.159	40.67s
TimesNet	0.678	0.554	0.527	7.94s	0.181	0.183	0.014	42.10s	0.285	0.359	0.233	106.95s

5 Conclusion

In this work, we investigate whether complex architectures or specialized inductive biases are necessary for reconstruction-based anomaly detection. We treat multivariate time series as sequences of multi-channel patches and use a Transformer to learn global representations. Our results show that complex representation learning designs are not always necessary, and a simple architecture can achieve competitive performance for reconstruction-based anomaly detection. This finding does not imply that local features are unimportant; they remain useful, especially in scenarios where point anomalies dominate. We further evaluate its performance under transfer settings, as well as different model and dataset scaling settings.

Moreover, investigating whether minimal yet effective architectures exist in other anomaly detection paradigms, such as representation-based and one-class approaches, is an interesting direction for future research. Although model scaling does not consistently improve performance in our experiments, this phenomenon may be related to the architecture design or the characteristics of time-series. We leave further investigation of this issue for future work.

AI use statement

In this work, we primarily used generative AI tools for grammar correction, formatting, and language refinement. We did not use generative AI tools for developing the methodology, designing the experiments, conducting literature surveys, or generating research ideas. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Ahmed Abdulaal, Zhuanghua Liu, and Tomer Lancewicki. Practical approach to asynchronous multivariate time series anomaly detection and localization. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pp. 2485–2494, 2021.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 3395–3404, 2020.

Markus M Breunig, Hans-Peter Kriegel, Raymond T Ng, and Jörg Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pp. 93–104, 2000.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Liang Dai, Tao Lin, Chang Liu, Bo Jiang, Yanwei Liu, Zhen Xu, and Zhi-Li Zhang. Sdfvae: Static and dynamic factorized vae for anomaly detection of multivariate cdn kpis. In *Proceedings of the web conference 2021*, pp. 3076–3086, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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)*, pp. 4171–4186, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Emadeldeen Eldele, Mohamed Ragab, Zhenghua Chen, Min Wu, Chee Keong Kwoh, Xiaoli Li, and Cuntai Guan. Time-series representation learning via temporal and contextual contrasting. *arXiv preprint arXiv:2106.14112*, 2021.

Yuchen Fang, Jiandong Xie, Yan Zhao, Lu Chen, Yunjun Gao, and Kai Zheng. Temporal-frequency masked autoencoders for time series anomaly detection. In *2024 IEEE 40th international conference on data engineering (ICDE)*, pp. 1228–1241. IEEE, 2024.

Yuye Feng, Wei Zhang, Yao Fu, Weihao Jiang, Jiang Zhu, and Wenqi Ren. Sensitivehue: Multivariate time series anomaly detection by enhancing the sensitivity to normal patterns. In *Proceedings of the 30th ACM SIGKDD Conference on knowledge discovery and data mining*, pp. 782–793, 2024.

Markus Goldstein and Andreas Dengel. Histogram-based outlier score (hbos): A fast unsupervised anomaly detection algorithm. *KI-2012: poster and demo track*, 1:59–63, 2012.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Tao Huang, Pengfei Chen, and Ruipeng Li. A semi-supervised vae based active anomaly detection framework in multivariate time series for online systems. In *Proceedings of the ACM web conference 2022*, pp. 1797–1806, 2022.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International conference on learning representations*, 2021.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Alexander Lavin and Subutai Ahmad. Evaluating real-time anomaly detection algorithms–the numenta anomaly benchmark. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*, pp. 38–44. IEEE, 2015.

Mengxuan Li, Ke Liu, Hongyang Chen, Jiajun Bu, Hongwei Wang, and Haishuai Wang. Tsinr: capturing temporal continuity via implicit neural representations for time series anomaly detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 671–682, 2025.

Zeyan Li, Nengwen Zhao, Shenglin Zhang, Yongqian Sun, Pengfei Chen, Xidao Wen, Minghua Ma, and Dan Pei. Constructing large-scale real-world benchmark datasets for aiops. *arXiv preprint arXiv:2208.03938*, 2022.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. *Advances in Neural Information Processing Systems*, 37:108231–108261, 2024.

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

Youngeun Nam, Susik Yoon, Yooju Shin, Minyoung Bae, Hwanjun Song, Jae-Gil Lee, and Byung Suk Lee. Breaking the time-frequency granularity discrepancy in time-series anomaly detection. In *Proceedings of the ACM Web Conference 2024*, pp. 4204–4215, 2024.

Y Nie. A time series is worth 64words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

Kohei Obata, Yasuko Matsubara, and Yasushi Sakurai. Robust and explainable detector of time series anomaly via augmenting multiclass pseudo-anomalies. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 2198–2209, 2025.

John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S Tsay, Aaron J Elmore, and Michael J Franklin. Volume under the surface: A new accuracy evaluation measure for time-series anomaly detection. *Proc. VLDB Endow.*, 15(11):2774–2787, 2022.

Daehyung Park, Yuuna Hoshi, and Charles C Kemp. A multimodal anomaly detector for robot-assisted feeding using an lstm-based variational autoencoder. *IEEE Robotics and Automation Letters*, 3(3):1544–1551, 2018.

Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. Anomaly detection in time series: a comprehensive evaluation. 2022.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2828–2837, 2019.

Nesime Tatbul, Tae Jun Lee, Stan Zdonik, Mejbah Alam, and Justin Gottschlich. Precision and recall for time series. *Advances in neural information processing systems*, 31, 2018.

Shreshth Tuli, Giuliano Casale, and Nicholas R Jennings. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *arXiv preprint arXiv:2201.07284*, 2022.

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

Rui Wang, Xudong Mou, Renyu Yang, Kai Gao, Pin Liu, Chongwei Liu, Tianyu Wo, and Xudong Liu. Cutaddpaste: Time series anomaly detection by exploiting abnormal knowledge. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3176–3187, 2024.

Qingsong Wen, Jingkun Gao, Xiaomin Song, Liang Sun, Huan Xu, and Shenghuo Zhu. Robuststl: A robust seasonal-trend decomposition algorithm for long time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 5409–5416, 2019.

Qingsong Wen, Linxiao Yang, Tian Zhou, and Liang Sun. Robust time series analysis and applications: An industrial perspective. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 4836–4837, 2022.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*, 2022.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. Catch: Channel-aware multivariate time series anomaly detection via frequency patching. In *International conference on learning representations*, volume 2025, pp. 17017–17045, 2025.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv preprint arXiv:2110.02642*, 2021.

Mohammed A Yahya, Antonio R Moya, and Sebastián Ventura. Deep learning for multivariate time series anomaly detection: An evaluation of reconstruction-based methods. *Artificial Intelligence Review*, 58(12):400, 2025.

Zhihan Yue, Yujing Wang, Juanyong Duan, Tianmeng Yang, Congrui Huang, Yunhai Tong, and Bixiong Xu. Ts2vec: Towards universal representation of time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 8980–8987, 2022.

Zahra Zamanzadeh Darban, Geoffrey I Webb, Shirui Pan, Charu Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1):1–42, 2024.

Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, and Marinka Zitnik. Self-supervised contrastive pre-training for time series via time-frequency consistency. *Advances in neural information processing systems*, 35:3988–4003, 2022.

Zhijie Zhong, Zhiwen Yu, Xing Xi, Yue Xu, Wenming Cao, Yiyuan Yang, Kaixiang Yang, and Jane You. Simad: A simple dissimilarity-based approach for time-series anomaly detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

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

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.