CANDI: Curated Test-Time Adaptation for Multivariate Time-Series Anomaly Detection Under Distribution Shift

HyunGi Kim[1], Jisoo Mok[2], Hyungyu Lee[1], Juhyeon Shin[3], Sungroh Yoon[1, 3, 4 †]
[1]Department of ECE, Seoul National University
[2]DGIST
[3]IPAI, Seoul National University
[4]AIIS, ASRI, and INMC, Seoul National University
rlagusrl0128@snu.ac.kr, jmok@dgist.ac.kr, {rucy74, newjh12, sryoon}@snu.ac.kr

Abstract

Multivariate time-series anomaly detection (MTSAD) aims to identify deviations from normality in multivariate time-series and is critical in real-world applications. However, in real-world deployments, distribution shifts are ubiquitous and cause severe performance degradation in pre-trained anomaly detector. Test-time adaptation (TTA) updates a pre-trained model on-the-fly using only unlabeled test data, making it promising for addressing this challenge. In this study, we propose **CANDI** (**C**urated test-time adaptation for multivariate time-series **AN**omaly detection under **DI**stribution shift), a novel TTA framework that selectively identifies and adapts to potential false positives while preserving pre-trained knowledge. CANDI introduces a False Positive Mining (FPM) strategy to curate adaptation samples based on anomaly scores and latent similarity, and incorporates a plug-and-play Spatiotemporally-Aware Normality Adaptation (SANA) module for structurally informed model updates. Extensive experiments demonstrate that CANDI significantly improves the performance of MTSAD under distribution shift, improving AUROC up to 14% while using fewer adaptation samples.

Code — https://github.com/kimanki/CANDI

Introduction

Multivariate time-series anomaly detection (MTSAD) aims to identify abnormal patterns within multivariate time-series data, which contain multiple interdependent variables (Wang et al. 2025; Li and Jung 2023; Choi et al. 2021). This capability is essential for maintaining the stability, safety, and efficiency of complex real-world systems through continuous monitoring of their states and timely decision-making (Duan et al. 2024; Shin et al. 2020). Accurate and robust MTSAD models are thus crucial for ensuring the stable operation of high-stakes, time-sensitive applications, such as industrial maintenance (Tanuska et al. 2021) and healthcare monitoring (Galvão et al. 2024).

Due to the scarcity of labeled anomaly data in real-world scenarios, most MSTAD methods adopt unsupervised approaches (Xu et al. 2022; Song et al. 2023; Kim et al. 2025b). These methods generally assume that the training data consists only of normal operating conditions and learn to model normality. One common approach is reconstruction-based anomaly detection (Wu et al. 2025; Xu et al. 2022; Zhang et al. 2019). Here, the model, typically in the form of an autoencoder, is trained to reconstruct normal time-series data, and at test-time, samples that yield high reconstruction errors are identified as anomalies. Density-based methods estimate the probability density of normal time-series and flag data with low likelihoods as anomalies (Zhou et al. 2023; Dai and Chen 2022). Lastly, distance-based approaches detect anomalies by measuring the distance of test data to normal clusters in a learned embedding space (Shen, Li, and Kwok 2020; Kim et al. 2023).

Despite recent advances, most MTSAD approaches assume that the training and test data belong to the same distribution. However, this assumption is often violated in real-world systems, due to numerous factors that cause a shift in the data distribution, *e.g.,* changes in system dynamics, sensor drifts, or environmental changes (Karimi and Paul 2010). Such distribution shifts induce normality shifts, where previously unseen but normal patterns emerge in the

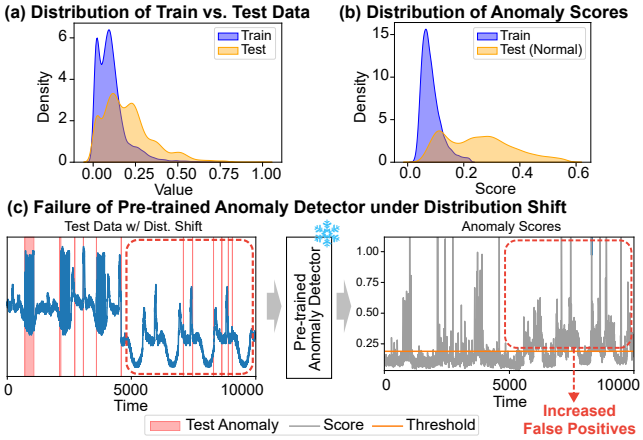

Figure 1: [Top] Real-world time-series data often exhibit non-stationarity, leading to continuous distribution shifts between training and test data. [Bottom] As shown in the later part of the anomaly scores, under distribution shift, pre-trained anomaly detectors can provide excessive false positives, undermining reliability under deployment.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.
† Corresponding Author

test data (Han et al. 2023). As shown in Figure 1, MTSAD models that have not been adapted to these shifts are prone to misclassifying these new normal patterns as anomalies, leading to a substantial increase in false positives.

Test-time adaptation (TTA) refers to the paradigm of adapting a pre-trained model at inference time using only unlabeled test data (Wang et al. 2020). While TTA has shown success in tasks such as image classification and segmentation (Lee et al. 2024; Gao, Yan, and He 2023), its application to MTSAD remains largely underexplored. In MTSAD, TTA offers a promising avenue towards addressing continuously evolving distribution. A prior work (Kim, Park, and Choo 2024) performs TTA on MTSAD by updating all trainable parameters on test samples that are identified as normal. This approach suffers from two major setbacks. First, it completely disregards false positives, *i.e.*, normal samples that are misclassified as anomalies. These false positives can provide informative learning signal as they correspond to underrepresented normal patterns, revealing regions where the model needs further adaptation. Second, adaptation of all trainable parameters may overwrite useful representations learned during pre-training.

To address the challenges of MTSAD posed by distribution shift, we propose **CANDI** (**C**urated test-time adaptation for multivariate time-series **AN**omaly detection under **DI**stribution shift), a novel TTA framework for MTSAD that adapts a pre-trained anomaly detector by curating informative test samples while preserving the original knowledge of the detector. CANDI is built on a reconstruction-based anomaly detector and introduces two key components: False Positive Mining (FPM) and Spatiotemporally-Aware Normality Adaptation (SANA). FPM identifies potential false positives based on their anomaly scores and proximity in latent space to normal validation samples. These challenging-to-detect yet reliable samples are used for adaptation.

Instead of updating the entire model, SANA provides a lightweight, plug-and-play adaptation module that captures temporal and inter-variable shifts via temporal convolutions and an attention mechanism, while keeping the backbone frozen. By combining selective adaptation signals with a safe adaptation mechanism, CANDI enhances robustness and accuracy under distribution shift without compromising the pre-trained model. Through extensive experiments, CANDI demonstrates significant gains over baselines under distribution shifts, including a 14% improvement of AUROC compared to the TTA baseline despite using less than 2% of the total test data for adaptation.

In summary, our contributions are as follows:

- We identify and address the critical challenge of distribution shift in MTSAD, a problem that causes significant false positives in real-world systems.
- We propose CANDI, a novel TTA framework for MTSAD that curates informative samples via false positive mining, and adapts the model with a spatiotemporally-aware module while preserving pre-trained knowledge.
- We demonstrate that CANDI consistently outperforms MTSAD baselines, achieving up to a 14% AUROC gain while using less than 2% of the data for adaptation.

Related Works

Unsupervised Multivariate Time-series Anomaly Detection

Unsupervised multivariate time-series anomaly detection (MTSAD) has been studied across diverse paradigms. Traditional methods like LOF (Breunig et al. 2000) and one-class SVM (Manevitz and Yousef 2001) have been applied, but often fail to capture temporal dependencies. Recent deep models fall into reconstruction-, density-, and graph-based categories. Reconstruction-based models detect anomalies based on reconstruction errors (Wu et al. 2025; Xu et al. 2022). USAD (Audibert et al. 2020) extends this approach by introducing adversarial training between dual decoders. Density-based models such as OmniAnomaly (Su et al. 2019) use variational autoencoders to detect low-likelihood patterns. Structure-aware models focus on inter-variable relations: MSCRED (Zhang et al. 2019) reconstructs multi-scale correlation maps, GDN (Deng and Hooi 2021) applies graph neural networks, and TimesNet (Wu et al. 2022) leverages frequency-aware blocks for improved detection.

However, most unsupervised MTSAD models assume a static normal distribution after training. In reality, normality may drift due to system aging, sensor noise, or environmental changes (Zhu et al. 2023; Liu et al. 2023; Han et al. 2023). Without adaptation, false positives increase over time. Our work addresses this by enabling test-time refinement using unlabeled but selectively informative samples.

Test-time Adaptation

Test-time adaptation (TTA) (Liang, He, and Tan 2025; Wang et al. 2020; Niu et al. 2023; Jia et al. 2024) is a paradigm that updates a pre-trained model at inference time to address distribution shifts, using only unlabeled test data. In the domain of image classification, TTA methods, such as TENT (Wang et al. 2020) and MEMO (Zhang, Levine, and Finn 2022), adjust model parameters by minimizing prediction entropy or self-supervised losses. More recent frameworks like CoTTA (Wang et al. 2022) further explore the continuous adaptation of a pre-trained image classifier while addressing the risk of catastrophic forgetting.

Due to the ever-evolving, dynamic nature of real-world time-series data, extending TTA to time-series data is a natural yet under-explored direction (Kim et al. 2025a). TTA allows models to track evolving normal distribution and maintain performance under non-stationary conditions. For instance, M2N2 (Kim, Park, and Choo 2024) proposes to adaptively detrend input signals and update model parameters using test samples predicted as normal, demonstrating the potential of TTA in MTSAD.

However, the existing TTA approach for MTSAD considers normality shift narrowly, focusing primarily on changes in overall trends while overlooking more complex temporal and inter-variable distribution shifts. Their reliance on limited adaptation cues and the practice of updating the full model can increase the risk of catastrophic forgetting by overwriting pre-trained knowledge. Furthermore, they overlook the adaptation potential of difficult-to-detect yet informative false positive samples.

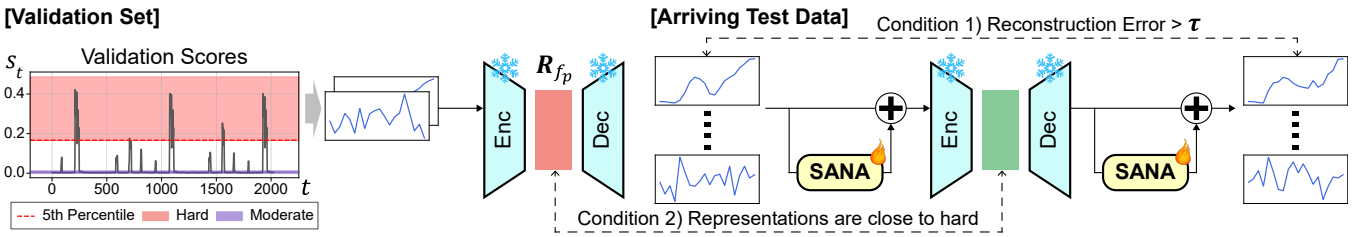

Figure 2: Overall framework of CANDI. [Left] Anomaly scores are first computed on a normal validation set, and latent representations of samples falling within the top α-percentile (*e.g.*, 5th percentile) are extracted and stored in a reference false positive set $\mathcal{R}_{fp}$. [Right] For arriving test data, if the anomaly score is above the threshold, its latent representation is compared to those in $\mathcal{R}_{fp}$. If the distance is sufficiently small, the sample is identified as a potential false positive and used for adaptation. Adaptation is performed via the plug-and-play *Spatiotemporally-Aware Normality Adaptation* (SANA) module, which updates only a lightweight residual component while preserving the knowledge and latent space of the pre-trained anomaly detector.

CANDI: Curated Test-time Adaptation for Multivariate Time-series Anomaly Detection

In this section, we present **CANDI**, a TTA framework for MTSAD. As illustrated in Figure 2, CANDI selectively adapts a pre-trained anomaly detector to curated informative test samples while preserving the pre-trained knowledge. It is comprised of two components: (1) *False Positive Mining* that selects potential false positives based on anomaly score distribution and similarity in the latent space, and (2) *Spatiotemporally-Aware Normality Adaptation* module that handles test-time distribution shifts in temporal and inter-variable patterns.

Problem Formulation

Let $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T] \in \mathbb{R}^{D \times T}$ denote a multivariate time-series with D variables over T time steps, where $\boldsymbol{x}_t \in \mathbb{R}^D$ is the observation at time t. The goal of MTSAD is to detect time steps or segments that deviate from the normal distribution. The train, validation, and test data are obtained by splitting $\boldsymbol{X}$ into contiguous segments in chronological order. Following the standard assumption in unsupervised MTSAD, the train and validation data consist solely of normal, while the test data include anomalies to be detected.

CANDI adopts a reconstruction-based approach, in which a pre-trained anomaly detector $f_{\boldsymbol{\theta}}$ is an autoencoder trained to reconstruct a sliding input window of length L:

$$\hat{\boldsymbol{X}}_{t-L+1:t} = f_{\boldsymbol{\theta}}(\boldsymbol{X}_{t-L+1:t}). \tag{1}$$

At inference time, the anomaly score s_t is computed as the average squared reconstruction error:

$$s_t = \frac{1}{D \cdot L} \left\| \boldsymbol{X}_{t-L+1:t} - \hat{\boldsymbol{X}}_{t-L+1:t} \right\|_2^2, \tag{2}$$

where higher scores indicate potential anomalies. However, in real-world deployments, the test distribution may shift due to factors like sensor drift or changing system dynamics. Our goal is to adapt the model to such shifts at test-time, without relying on labeled data or retraining the full model.

False Positive Mining

Rather than adapting to all test samples with low anomaly scores, we selectively identify samples that cannot easily be detected and thus are likely to contribute meaningfully to adaptation. This curated sample selection not only mitigates the risk of performance degradation by avoiding unreliable samples for adaptation, but also improves adaptation efficiency by reducing the number of test samples used. In particular, we focus on samples that are challenging for the model to detect, such as potential false positives that reflect ambiguous or underrepresented normality. In consequence, CANDI focuses on areas where the model's prediction is uncertain, improving robustness with fewer updates.

We first compute anomaly scores for all samples in a validation set that contains only normal data. Let $\boldsymbol{\mathcal{S}}_{\text{val}} = \{s_i^{\text{val}}\}_{i=1}^{N_{\text{val}}}$ denote the set of anomaly scores computed on this validation set, where s_i^{val} is the anomaly score for the i-th validation sample and N_{val} is the total number of samples in the validation set. We then define a threshold τ as the α-percentile of this score set:

$$\tau = \text{Percentile}(\boldsymbol{\mathcal{S}}_{\text{val}}, \alpha). \tag{3}$$

Following standard practice, test samples with $s_t > \tau$ are initially considered to be anomalous. However, since some normal samples in the validation set also exceed the same threshold τ (*i.e.*, $s_i^{\text{val}} > \tau$), we hypothesize that a subset of high-scoring test samples may likewise be false positives—normal but difficult instances that the model failed to capture during training. To identify these false positives, we collect validation samples with $s_i^{\text{val}} > \tau$ and extract their latent representations using the frozen pre-trained encoder: $\boldsymbol{z}_i = f^{\text{enc}}(\boldsymbol{X}_i^{\text{val}})$. We aggregate these into a reference set of false positive samples, denoted by $\mathcal{R}_{\text{fp}}$.

These reference samples reflect normal instances that exhibit unexpectedly high anomaly scores, suggesting that they lie near the decision boundary and share latent features with difficult-to-classify cases. To identify potential test-time false positives, we measure their proximity to known false positives from the validation set in the latent space using Mahalanobis distance. To provide stability, we estimate the mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ from the full set of latent representations of validation samples:

$$\boldsymbol{\mu}, \boldsymbol{\Sigma} = \text{MeanCov}(\mathcal{R}_{\text{val}}), \quad \mathcal{R}_{\text{val}} = \{f^{\text{enc}}(\boldsymbol{X}_i^{\text{val}})\}_{i=1}^{N_{\text{val}}}. \tag{4}$$

Using the full normal validation set ensures that the latent distance is measured with respect to the overall distribution

of normal patterns, providing robustness and avoiding bias from sparsely sampled or ambiguous subsets like $\mathcal{R}_{\text{fp}}$.

For each test sample predicted as anomalous ($s_t > \tau$), we compute its latent representation $\boldsymbol{z}_t = f^{\text{enc}}(\boldsymbol{X}_{t-L+1:t})$ and calculate its minimum squared Mahalanobis distance to the reference set $\mathcal{R}_{\text{fp}}$:

$$\mathcal{D}_M^2(\boldsymbol{z}_t, \mathcal{R}_{\text{fp}}) = \min_{\boldsymbol{z}_r \in \mathcal{R}_{\text{fp}}} (\boldsymbol{z}_t - \boldsymbol{z}_r)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}_t - \boldsymbol{z}_r). \quad (5)$$

We consider $\boldsymbol{X}_{t-L+1:t}$ a potential false positive and include it for adaptation if this distance is below the 5th percentile of the chi-squared distribution with latent dimension d. The threshold is defined as:

$$\delta = F^{-1}_{\chi^2_d}(0.05), \quad (6)$$

where $F^{-1}_{\chi^2_d}(\cdot)$ denotes the inverse cumulative distribution function. The inclusion criterion becomes:

$$\mathcal{D}_M^2(\boldsymbol{z}_t, \mathcal{R}_{\text{fp}}) < \delta. \quad (7)$$

This thresholding strategy is grounded in the statistical property that the squared Mahalanobis distance follows a chi-squared distribution with d degrees of freedom when latent representations of normal samples approximately follow a multivariate Gaussian. Thus, δ defines a tight neighborhood around the reference set in latent space, and selecting the 5th percentile ensures only test samples with representations sufficiently close to those of high-scoring validation samples are selected. These samples likely share subtle but informative patterns, making them suitable candidates for adaptation. The adaptation set $\mathcal{A}$ is constructed as:

$$\mathcal{A} = \left\{ \boldsymbol{X}_{t-L+1:t} \,\middle|\, (s_t > \tau) \wedge \left(\mathcal{D}_M^2\left(\boldsymbol{z}_t, \mathcal{R}_{\text{fp}}\right) < \delta\right) \right\}. \quad (8)$$

To further improve robustness, we also incorporate predicted normal samples with moderately high anomaly scores into the adaptation process. Specifically, we identify validation samples whose scores fall within the interquartile range (Q_1–Q_3) and store their latent representations as a separate reference set $\mathcal{R}_{\text{mod}}$. These samples are not clearly anomalous but deviate enough from typical patterns to indicate areas where the model's understanding of normality may be incomplete. For each test sample whose anomaly score is smaller than the threshold τ, we compute its squared Mahalanobis distance to $\mathcal{R}_{\text{mod}}$ using the same criterion as before. If the distance falls below the threshold δ, the sample is included in the final adaptation set $\mathcal{A}$.

Spatiotemporally-Aware Normality Adaptation

To enable stable and efficient TTA while preserving the knowledge of a pre-trained detector, we introduce a lightweight plug-and-play normality adaptation module, as illustrated in Figure 3. Motivated by TAFAS (Kim et al. 2025a), the module is attached to both the input and output of a frozen reconstruction-based anomaly detector. However, unlike TAFAS, which uses independent per-variable simple linear layers, we design a spatiotemporally-aware module composed of temporal convolution and inter-variable attention (Liu et al. 2024). This allows our approach

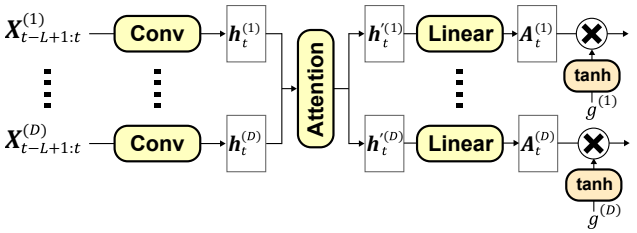

Figure 3: Architecture of the Spatiotemporally-Aware Normality Adaptation (SANA) module.

to capture distributional shifts occurring not only within each variable but also across variables through their interactions, providing a more expressive adaptation.

The input-side normality adaptation module adjusts incoming test samples to better align with the training-time normality distribution, enabling the pre-trained detector to process them effectively. Conversely, the output-side normality adaptation module transforms the reconstruction results to match the test-time normal distribution, compensating for any residual shift. This design preserves a consistent latent space, which is crucial for reliable false positive mining, while allowing flexible adaptation.

Input Normality Adaptation. Given a multivariate input window $\boldsymbol{X}_{t-L+1:t} \in \mathbb{R}^{D \times L}$, we first model the temporal dynamics of each variable independently. For the i-th variable, the univariate sequence $\boldsymbol{X}^{(i)}_{t-L+1:t} \in \mathbb{R}^L$ is encoded via a 1D convolution layer:

$$\boldsymbol{h}_t^{(i)} = \text{Conv}^{\text{in}}(\boldsymbol{X}^{(i)}_{t-L+1:t}). \quad (9)$$

The resulting temporal embeddings $\{\boldsymbol{h}_t^{(i)}\}_{i=1}^D$ are then processed by a single inter-variable attention layer to capture cross-variable dependencies:

$$\{\boldsymbol{h}_t'^{(i)}\}_{i=1}^D = \text{ATTN}^{\text{in}}(\{\boldsymbol{h}_t^{(i)}\}_{i=1}^D). \quad (10)$$

We apply a variable-wise linear layer to the attention output to compute the adjustment term $\boldsymbol{A}_t^{(i)} \in \mathbb{R}^L$ for each variable:

$$\boldsymbol{A}_t^{(i)} = \text{Linear}^{\text{in}}(\boldsymbol{h}_t'^{(i)}). \quad (11)$$

Finally, a learnable gating parameter $g^{(i)}$, activated by a $\tanh$ function, is used to modulate the adjustment applied to each variable:

$$\tilde{\boldsymbol{X}}^{(i)}_{t-L+1:t} = \boldsymbol{X}^{(i)}_{t-L+1:t} + \tanh(g^{(i)}) \cdot \boldsymbol{A}_t^{(i)}. \quad (12)$$

The adapted input $\tilde{\boldsymbol{X}}_{t-L+1:t}$ is then passed to the frozen pre-trained anomaly detector f_θ to obtain the reconstruction:

$$\hat{\boldsymbol{X}}_{t-L+1:t} = f_\theta(\tilde{\boldsymbol{X}}_{t-L+1:t}). \quad (13)$$

Output Normality Adaptation. To account for distributional shifts in the reconstructed space, we apply an output normality adaptation module with the same architectural structure. Each variable-wise reconstructed sequence $\hat{\boldsymbol{X}}^{(i)}_{t-L+1:t} \in \mathbb{R}^L$ is encoded using a 1D convolution layer:

$$\boldsymbol{h}_t^{(i,\text{out})} = \text{Conv}^{\text{out}}(\hat{\boldsymbol{X}}^{(i)}_{t-L+1:t}). \quad (14)$$

Dataset	Metric	$\alpha = 0.5\%$			$\alpha = 1.0\%$			$\alpha = 5.0\%$		
		Pretrained	M2N2	CANDI	Pretrained	M2N2	CANDI	Pretrained	M2N2	CANDI
SWaT	AUROC	0.827	**0.891**	0.889	0.827	0.891	**0.890**	0.827	0.891	**0.888**
	AUPRC	0.719	0.771	**0.779**	0.719	0.771	**0.780**	0.719	0.772	**0.781**
	F1	0.291	0.711	**0.752**	0.287	0.700	**0.738**	0.291	**0.636**	0.624
SMD_1-7	AUROC	0.883	0.901	**0.922**	0.883	0.864	**0.923**	0.883	0.886	**0.922**
	AUPRC	0.703	0.728	**0.736**	0.703	0.734	**0.737**	0.703	0.718	**0.737**
	F1	0.103	**0.662**	0.107	0.562	**0.707**	0.688	0.724	0.723	**0.725**
SMD_1-8	AUROC	0.719	0.837	**0.872**	0.719	0.805	**0.872**	0.719	0.772	**0.867**
	AUPRC	0.332	0.407	**0.432**	0.332	0.376	**0.434**	0.332	0.354	**0.423**
	F1	0.377	**0.406**	0.393	0.362	0.389	**0.409**	0.092	0.115	**0.213**
SMD_2-1	AUROC	0.648	0.698	**0.725**	0.648	0.693	**0.711**	0.648	0.683	**0.780**
	AUPRC	0.275	0.307	**0.319**	0.275	0.302	**0.314**	0.275	0.296	**0.348**
	F1	0.265	0.266	0.266	**0.296**	0.295	0.292	0.273	0.309	**0.327**
SMD_2-4	AUROC	0.821	0.895	**0.908**	0.821	0.895	**0.908**	0.821	0.828	**0.899**
	AUPRC	0.457	0.605	**0.608**	0.457	0.605	**0.608**	0.457	0.461	**0.600**
	F1	**0.357**	0.355	0.352	**0.378**	0.377	0.372	0.316	0.311	**0.512**
SMD_3-2	AUROC	0.451	0.573	**0.717**	0.451	0.632	**0.717**	0.451	0.640	**0.717**
	AUPRC	0.159	0.174	**0.199**	0.159	0.179	**0.199**	0.159	0.188	**0.199**
	F1	0.017	0.017	0.017	**0.031**	0.030	0.030	0.247	0.194	**0.262**

Table 1: Performance of test-time adaptation methods for multivariate time-series anomaly detection under test-time distribution shift. Bold denotes the best result for each metric and threshold level. Each threshold is determined by the α-percentile of validation anomaly scores.

The resulting temporal embeddings $\{\boldsymbol{h}_t^{(i,\text{out})}\}_{i=1}^{D}$ are passed through an inter-variable attention layer:

$$\{\boldsymbol{h}_t'^{(i,\text{out})}\}_{i=1}^{D} = \text{ATTN}^{\text{out}}(\{\boldsymbol{h}_t^{(i,\text{out})}\}_{i=1}^{D}). \quad (15)$$

Each attention output is then passed through a variable-wise linear layer to obtain the adjustment term:

$$\boldsymbol{A}_t^{(i,\text{out})} = \text{Linear}^{\text{out}}(\boldsymbol{h}_t'^{(i,\text{out})}). \quad (16)$$

Finally, a learnable gating parameter $g^{(i,\text{out})}$ modulates the adjustment via a $\tanh$ activation:

$$\tilde{\hat{\boldsymbol{X}}}_{t-L+1:t}^{(i)} = \hat{\boldsymbol{X}}_{t-L+1:t}^{(i)} + \tanh(g^{(i,\text{out})}) \cdot \boldsymbol{A}_t^{(i,\text{out})}. \quad (17)$$

Adaptation Objective. Only the parameters of the SANA modules are updated during test-time, while the pre-trained backbone parameters $\boldsymbol{\theta}$ remain frozen. We minimize the reconstruction loss over the selected adaptation set $\boldsymbol{\mathcal{A}}$:

$$\mathcal{L}_{\text{adapt}} = \frac{1}{D \cdot L} \sum_{\boldsymbol{X}_t \in \boldsymbol{\mathcal{A}}} \left\| \boldsymbol{X}_{t-L+1:t} - \tilde{\hat{\boldsymbol{X}}}_{t-L+1:t} \right\|_2^2. \quad (18)$$

This modular design enables the model to adapt to both temporal and relational distributional shifts at test-time, while preserving the generalization capabilities of the frozen backbone. By updating only lightweight modules with selectively chosen, informative test samples, our framework achieves robust and efficient test-time adaptation for anomaly detection.

Experiments

Experimental Setup

Datasets. We conduct experiments on representative multivariate time-series anomaly detection benchmarks: SWaT (Goh et al. 2016) and SMD (Su et al. 2019). SWaT is industrial control system datasets containing labeled normal and attack periods, reflecting real-world operational and environmental shifts. SMD is a dataset collected from server machines, organized into multiple subdatasets based on the entity. Since distribution shifts are not uniformly present across all SMD subsets, we select a representative subset of 5 server entities that exhibit prominent normality shifts: SMD_1-7, SMD_1-8, SMD_2-1, SMD_2-4 and SMD_3-2. We also evaluate CANDI on the 200 multivariate time-series datasets provided by the TSB-AD benchmark (Liu and Paparrizos 2024) to further assess robustness under a broader and more diverse collection of real-world conditions.

Baselines. To assess the benefits of test-time adaptation, we compare our method to M2N2 (Kim, Park, and Choo 2024), a recent approach that applies TTA to time-series anomaly detection. Following the original setup in M2N2, we use an MLP-based autoencoder as the pre-trained anomaly detector for both methods. For each dataset, we report the performance of: (1) the pre-trained model without adaptation, (2) the model adapted with M2N2, and (3) the model adapted with our proposed CANDI framework. This comparison allows us to isolate the effects of different adaptation strategies under the same model backbone.

Implementation Details. We evaluate detection performance using the standard metrics: AUROC and AUPRC. In addition, we report F1 scores at fixed false positive rate (FPR) thresholds determined by the α-percentile of validation anomaly scores. Specifically, we report results for $\alpha \in \{0.5\%, 1\%, 5\%\}$. The pre-trained models are trained using the Adam optimizer (Kingma and Ba 2015) with an ini-

FPM	SANA	SWaT			SMD_1-8			SMD_2-1			SMD_2-4			SMD_3-2		
		ROC	PRC	F1	ROC	PRC	F1	ROC	PRC	F1	ROC	PRC	F1	ROC	PRC	F1
w/o TTA		0.83	0.72	0.29	0.72	0.33	0.10	0.65	0.28	0.27	0.82	0.46	0.32	0.45	0.16	0.25
✗	✗	0.89	0.77	0.64	0.77	0.35	0.12	0.68	0.30	0.31	0.83	0.46	0.31	0.64	0.19	0.19
✓	✗	0.80	0.71	0.22	0.69	0.32	0.01	0.71	0.34	0.20	**0.93**	**0.66**	0.25	**0.74**	**0.22**	0.21
✗	✓	0.89	0.78	0.64	0.83	0.36	0.15	0.71	0.32	0.33	0.83	0.46	0.37	0.68	0.19	0.25
✓	✓	0.89	0.78	0.62	**0.87**	**0.42**	**0.21**	**0.78**	**0.35**	0.33	0.90	0.60	**0.51**	0.72	0.20	**0.26**

Table 2: Ablation study of CANDI across five datasets at $\alpha = 5.0\%$. **FPM** and **SANA** denote False Positive Mining and Spatiotemporally-Aware Normality Adaptation, respectively. ROC and PRC denote AUROC and AUPRC, respectively.

tial learning rate of 0.001 and cosine learning rate scheduling (Loshchilov and Hutter 2017) for 30 epochs. M2N2 is re-implemented to match our experimental setup for fair comparison. All experiments are conducted using three different random seeds, and we report the average performance. Full results, including standard deviations and additional implementation details, are included in the Appendix.

Evaluating Test-time Adaptation in Multivariate Time-series Anomaly Detection

Table 1 compares the performance of the pre-trained anomaly detector, M2N2, and CANDI under test-time distribution shift on multiple multivariate time-series anomaly detection benchmarks. We report AUROC, AUPRC, and F1 scores across three anomaly score thresholds, $\alpha \in \{0.5\%, 1.0\%, 5.0\%\}$, with τ set as the α-percentile of validation scores. Smaller α yields stricter thresholds with fewer false positives, while larger α reflects more relaxed thresholds, admitting more ambiguous cases.

CANDI consistently matches or outperforms both baselines, with the largest gains seen where the pre-trained model struggles. For instance, on SMD_1-8, CANDI improves AUROC from 0.719 (pre-trained) and 0.772 (M2N2) to 0.867 at $\alpha = 5.0\%$. Similarly, on SMD_3-2, AUROC rises from 0.451 to 0.717—a 59.0% relative improvement—showing CANDI's strength in challenging conditions.

Notably, at $\alpha = 5.0\%$, where many false positives emerge due to a lower threshold, CANDI turns this challenge into an advantage. The false positive mining selects informative samples close to trusted normal patterns in latent space and adapts using the lightweight SANA module. As a result, CANDI achieves substantial gains; for example, on SMD_2-4, F1 improves from 0.316 (pre-trained) and 0.311 (M2N2) to 0.512, with AUPRC increasing to 0.600. We provide further evaluation results on the large-scale TSB-AD benchmark (Liu and Paparrizos 2024) in the Appendix.

Ablation Study

Table 2 presents an ablation study evaluating the contributions of the two core components of CANDI: False Positive Mining (FPM) and Spatiotemporally-Aware Normality Adaptation (SANA), across five datasets at $\alpha = 5.0\%$. When FPM is disabled, the model performs adaptation using all test samples whose anomaly scores fall below the threshold. When SANA is disabled, entire parameters of the pre-trained anomaly detector are updated during adaptation. When both FPM and SANA are disabled, the model corresponds to M2N2, which adapts the entire pre-trained anomaly detector using all test samples whose anomaly scores fall below the threshold. This setting serves as a baseline for assessing the effectiveness of each component.

When FPM is used without SANA, we observe performance degradation across several datasets. For example, on SWaT, F1 drops from 0.64 (M2N2) to 0.22, and on SMD_2-1, from 0.31 to 0.20. This suggests that although FPM improves the adaptation sample quality by mining informative candidates, the adaptation process without the structural constraint of SANA updates distorts the pre-trained latent space, undermining the reliability of FPM's latent similarity calculations. These results underscore the necessity of freezing the pre-trained model and adapting only the SANA module to preserve latent consistency.

In contrast, using only SANA without FPM leads to consistent gains over M2N2. For instance, on SMD_2-4, F1 improves from 0.31 to 0.37, and on SMD_1-8, from 0.12 to 0.15. This shows that even without selective sample mining, restricting updates to a lightweight, structurally-informed module like SANA prevents catastrophic forgetting and enables stable test-time adaptation. However, this configuration does not leverage the informative signals present in potential false positives, which limits its ability to fully recover useful patterns missed during training.

The proposed CANDI framework, with both FPM and SANA enabled, achieves the best performance across the majority of datasets despite using fewer adaptation samples than M2N2. These results demonstrate that combining sample selection through FPM with the localized and structured updates enabled by SANA provides a robust and efficient adaptation strategy. By leveraging latent-consistent false positives while preserving the integrity of the pre-trained detector, CANDI achieves both superior accuracy and stable performance under distribution shift.

Analysis on ROC and PR Curves. Figure 4 shows the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves to further compare the detection capabilities. ROC curves show that CANDI consistently achieves a significantly higher true positive rate (TPR) at low false positive rates (*e.g.*, FPR = 0.1) compared to all baselines. This indicates that CANDI is more effective at identifying true anomalies while keeping false alarms low—a critical property in real-world deployment scenarios.

We also evaluate CANDI Hard, a variant that uses only the hard samples identified through false positive mining,

Dataset	Method	Mod. Samples	Ano. in Mod.	Hard Samples	Ano. in Hard	Total Adapt	Total Test	Total Ano.	ROC	PRC	F1
SMD_1-8	M2N2	N/A	N/A	N/A	N/A	11,760	23,690	943	0.772	0.354	0.115
	CANDI Hard	N/A	N/A	5,437	527	5,437			0.861	0.416	0.166
	CANDI Mod.	11,038	114	N/A	N/A	11,038			0.826	0.362	0.142
	CANDI (Hard + Mod.)	14,924	178	4,332	497	19,256			**0.867**	**0.423**	**0.213**
SMD_2-1	M2N2	N/A	N/A	N/A	N/A	22,573	23,685	1,287	0.683	0.296	0.309
	CANDI Hard	N/A	N/A	353	78	353			0.735	0.323	0.262
	CANDI Mod.	13,940	184	N/A	N/A	13,940			0.717	0.317	0.328
	CANDI (Hard + Mod.)	10,948	107	310	76	11,258			**0.780**	**0.348**	**0.327**

Table 3: Comparison of different adaptation sample configurations on SMD_1-8 and SMD_2-1. "Mod." and "Hard" refer to the difficulty levels of the samples used. "Ano." refers to anomaly. "Total Adapt" denotes the number of adaptation samples.

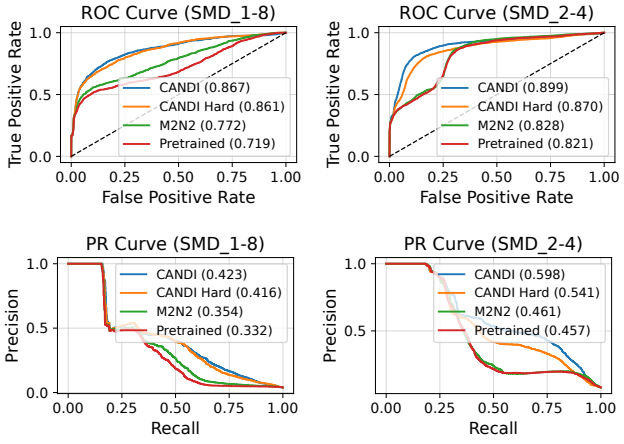

Figure 4: Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curves for anomaly detection. The value in parentheses indicates the area under each curve.

excluding moderate samples. Despite using fewer adaptation samples, this variant still outperforms the baselines and achieves performance close to the original CANDI. These results highlight that even partial adaptation guided by carefully curated samples can offer substantial benefits, and that the CANDI framework further enhances performance by incorporating additional reliable test-time samples.

Comparison of Adaptation Sample Configurations. Table 3 compares the effectiveness of different adaptation sample configurations, including the proposed CANDI variants and M2N2. Notably, CANDI with only hard samples outperforms M2N2 while using significantly fewer adaptation samples—less than half on SMD_1-8 (5,437 vs. 11,760) and less than 2% on SMD_2-1 (353 vs. 22,573). Despite this drastic reduction, it achieves superior AUROC and AUPRC, highlighting the efficiency of our curated adaptation.

Among the samples identified as potential false positives for adaptation, some fraction are actual anomalies: approximately 10% on SMD_1-8 and 25% on SMD_2-1. This indicates that CANDI sometimes performs adaptation on mislabeled anomalous data. Nevertheless, it still outperforms the baseline, suggesting robustness to moderate contamination. These results highlight that enhancing the accuracy of false positive mining, thereby reducing the potential negative impact of anomaly adaptation, is an important future direction for TTA frameworks in MTSAD.

Method	SMD_1-8 ROC	SMD_1-8 PRC	SMD_1-8 F1	SMD_2-1 ROC	SMD_2-1 PRC	SMD_2-1 F1
Linear	0.840	0.415	0.115	0.673	0.294	0.275
SANA	0.867	0.423	0.213	0.780	0.348	0.327

Table 4: Comparison between Linear and SANA adaptation modules on SMD_1-8 and SMD_2-1. ROC and PRC denote AUROC and AUPRC, respectively.

Among all configurations, the best performance is achieved when both moderate and hard samples are used together. The results demonstrate that combining reliable moderate samples with carefully mined hard samples enables more comprehensive and effective test-time adaptation under distribution shift.

Effectiveness of SANA Architecture. Table 4 compares the proposed SANA module with a linear adaptation head on SMD_1-8 and SMD_2-1. Across all metrics, SANA outperforms the linear approach—achieving notably higher F1 scores (0.213 vs. 0.115 on SMD_1-8 and 0.327 vs. 0.275 on SMD_2-1). This indicates that SANA provides more effective test-time adaptation. The performance gap highlights the importance of structure-aware adaptation. Unlike linear updates, SANA captures temporal and variable-wise dependencies while preserving the pre-trained model's latent space. This allows SANA to adapt meaningfully under distribution shifts without degrading the original detector.

Conclusion

Multivariate time-series anomaly detection in deployment environments suffers from performance degradation due to distribution shift. To address this, we proposed **CANDI**, a test-time adaptation framework that curates informative false positives and adapts using a lightweight, structure-aware module. CANDI combines *False Positive Mining* (FPM) to identify reliable adaptation samples with *Spatiotemporally-Aware Normality Adaptation* (SANA), a plug-and-play module that preserves pre-trained knowledge. Experiments show that CANDI significantly outperforms prior methods, especially under relaxed anomaly thresholds.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [No.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University); No.RS-2024-00357879, AI-based Biosignal Fusion and Generation Technology for Intelligent Personalized Chronic Disease Management], the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2022R1A3B1077720; 2022R1A5A708390811; No.RS-2023-00212484, xAI for Motion Prediction in Complex, Real-World Driving Environment), the BK21 FOUR program of the Education and the Research Program for Future ICT Pioneers, Seoul National University in 2025, Hyundai Motor Company, Samsung Electronics Co., Ltd (IO250624-13143-01), and HUINNO AIM Company through HA-Rnd2325-predictClinicalDeterioration.

References

Audibert, J.; Michiardi, P.; Guyard, F.; Marti, S.; and Zuluaga, M. A. 2020. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, 3395–3404.

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; and Sander, J. 2000. LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, 93–104.

Choi, K.; Yi, J.; Park, C.; and Yoon, S. 2021. Deep learning for anomaly detection in time-series data: Review, analysis, and guidelines. *IEEE access*, 9: 120043–120065.

Dai, E.; and Chen, J. 2022. Graph-Augmented Normalizing Flows for Anomaly Detection of Multiple Time Series. In *International Conference on Learning Representations*.

Deng, A.; and Hooi, B. 2021. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 4027–4035.

Duan, Y.; Xue, K.; Sun, H.; Bao, H.; Wei, Y.; You, Z.; Zhang, Y.; Jiang, X.; Yang, S.; Chen, J.; Duan, B.; and Ou, Z. 2024. LogEDL: Log Anomaly Detection via Evidential Deep Learning. *Applied Sciences*, 14(16).

Galvão, Y. M.; Castro, L.; Ferreira, J.; Neto, F. B. d. L.; Fagundes, R. A. d. A.; and Fernandes, B. J. 2024. Anomaly detection in smart houses for healthcare: Recent advances, and future perspectives. *SN Computer Science*, 5(1): 136.

Gao, Z.; Yan, S.; and He, X. 2023. ATTA: Anomaly-aware Test-Time Adaptation for Out-of-Distribution Detection in Segmentation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Goh, J.; Adepu, S.; Junejo, K. N.; and Mathur, A. 2016. A dataset to support research in the design of secure water treatment systems. In *International conference on critical information infrastructures security*, 88–99. Springer.

Han, D.; Wang, Z.; Chen, W.; Wang, K.; Yu, R.; Wang, S.; Zhang, H.; Wang, Z.; Jin, M.; Yang, J.; et al. 2023. Anomaly Detection in the Open World: Normality Shift Detection, Explanation, and Adaptation. In *NDSS*.

Jia, H.; Kwon, Y.; Orsino, A.; Dang, T.; Talia, D.; and Mascolo, C. 2024. TinyTTA: Efficient Test-time Adaptation via Early-exit Ensembles on Edge Devices. *Advances in Neural Information Processing Systems*, 37: 43274–43299.

Karimi, A.; and Paul, M. R. 2010. Extensive chaos in the Lorenz-96 model. *Chaos: An interdisciplinary journal of nonlinear science*, 20(4).

Kim, D.; Park, S.; and Choo, J. 2024. When model meets new normals: Test-time adaptation for unsupervised time-series anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 13113–13121.

Kim, H.; Kim, S.; Min, S.; and Lee, B. 2023. Contrastive Time-Series Anomaly Detection. *IEEE Transactions on Knowledge and Data Engineering*.

Kim, H.; Kim, S.; Mok, J.; and Yoon, S. 2025a. Battling the non-stationarity in time series forecasting via test-time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 17868–17876.

Kim, H.; Mok, J.; Lee, D.; Lew, J.; Kim, S.; and Yoon, S. 2025b. Causality-Aware Contrastive Learning for Robust Multivariate Time-Series Anomaly Detection. In *Forty-second International Conference on Machine Learning*.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Lee, J.; Jung, D.; Lee, S.; Park, J.; Shin, J.; Hwang, U.; and Yoon, S. 2024. Entropy is not Enough for Test-Time Adaptation: From the Perspective of Disentangled Factors. In *The Twelfth International Conference on Learning Representations*.

Li, G.; and Jung, J. J. 2023. Deep learning for anomaly detection in multivariate time series: Approaches, applications, and challenges. *Information Fusion*, 91: 93–102.

Liang, J.; He, R.; and Tan, T. 2025. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1): 31–64.

Liu, J.; Yang, D.; Zhang, K.; Gao, H.; and Li, J. 2023. Anomaly and change point detection for time series with concept drift. *World Wide Web*, 26(5): 3229–3252.

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

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2024. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.

Loshchilov, I.; and Hutter, F. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon,*

France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net.

Manevitz, L. M.; and Yousef, M. 2001. One-class SVMs for document classification. *Journal of machine Learning research*, 2(Dec): 139–154.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards Stable Test-time Adaptation in Dynamic Wild World. In *The Eleventh International Conference on Learning Representations*.

Shen, L.; Li, Z.; and Kwok, J. 2020. Timeseries Anomaly Detection using Temporal Hierarchical One-Class Network. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 13016–13026. Curran Associates, Inc.

Shin, Y.; Lee, S.; Tariq, S.; Lee, M. S.; Jung, O.; Chung, D.; and Woo, S. S. 2020. Itad: integrative tensor-based anomaly detection system for reducing false positives of satellite systems. In *Proceedings of the 29th ACM international conference on information & knowledge management*, 2733–2740.

Song, J.; Kim, K.; Oh, J.; and Cho, S. 2023. MEMTO: Memory-guided Transformer for Multivariate Time Series Anomaly Detection. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 57947–57963. Curran Associates, Inc.

Su, Y.; Zhao, Y.; Niu, C.; Liu, R.; Sun, W.; and Pei, D. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2828–2837.

Tanuska, P.; Spendla, L.; Kebisek, M.; Duris, R.; and Stremy, M. 2021. Smart anomaly detection and prediction for assembly process maintenance in compliance with industry 4.0. *Sensors*, 21(7): 2376.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2020. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*.

Wang, F.; Jiang, Y.; Zhang, R.; Wei, A.; Xie, J.; and Pang, X. 2025. A Survey of Deep Anomaly Detection in Multivariate Time Series: Taxonomy, Applications, and Directions. *Sensors (Basel, Switzerland)*, 25(1): 190.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7201–7211.

Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; and Long, M. 2022. Timesnet: Temporal 2d-variation modeling for general time series analysis. *arXiv preprint arXiv:2210.02186*.

Wu, X.; Qiu, X.; Li, Z.; Wang, Y.; Hu, J.; Guo, C.; Xiong, H.; and Yang, B. 2025. CATCH: Channel-Aware Multivariate Time Series Anomaly Detection via Frequency Patching. In *The Thirteenth International Conference on Learning Representations*.

Xu, J.; Wu, H.; Wang, J.; and Long, M. 2022. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. In *International Conference on Learning Representations*.

Zhang, C.; Song, D.; Chen, Y.; Feng, X.; Lumezanu, C.; Cheng, W.; Ni, J.; Zong, B.; Chen, H.; and Chawla, N. V. 2019. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 1409–1416.

Zhang, M.; Levine, S.; and Finn, C. 2022. Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems*, 35: 38629–38642.

Zhou, Q.; Chen, J.; Liu, H.; He, S.; and Meng, W. 2023. Detecting Multivariate Time Series Anomalies with Zero Known Label. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(4): 4963–4971.

Zhu, J.; Cai, S.; Deng, F.; Ooi, B. C.; and Zhang, W. 2023. METER: A Dynamic Concept Adaptation Framework for Online Anomaly Detection. *Proc. VLDB Endow.*, 17(4): 794–807.