DoKnowAD: Calibrating Normal Representations with Refined Domain Knowledge to Enhance Time Series Anomaly Detection

Shiwang Xing[1], Jianwei Niu[1,2,3], Tao Ren[4*]

[1]State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China.

[2] Hangzhou Innovation Institute of Beihang University, Zhejiang Key Laboratory of Industrial Big Data and Robot Intelligent Systems, Hangzhou, China.

[3]Zhengzhou University Research Institute of Industrial Technology, Zhengzhou University, Zhengzhou, China.

[4]Institute of Software Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing, China.

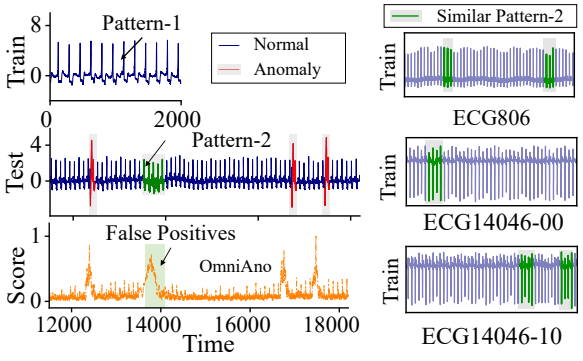
{xingshiwang, niujianwei}@buaa.edu.cn, rentao22@iscas.ac.cn

Abstract

Time series anomaly detection (TSAD) is critical in various real-world applications. Due to the high cost of manual annotation, unsupervised methods are commonly employed to distinguish abnormal patterns from normal ones based on data or representation characteristics. However, the limited coverage of a single dataset often leads to misclassifying test-time normal patterns that deviate from the training distribution as anomalies. In view of this, we propose to introduce domain knowledge from auxiliary datasets (AuxSets) to enhance domain-level normality understanding in the target dataset (TargetSet). However, through in-depth analysis on the representation space of the TargetSet after incorporating AuxSets, we find that consistent knowledge about normality from homogeneous AuxSets do little help to TargetSet, while diverse knowledge from heterogeneous AuxSets can bring semantic confusion of normality for TargetSet, both of which can degrade TargetSet detection performance. To address the issue, we design DoKnowAD, a framework that introduces a Representation HyperVolume Estimation metric to identify helpful heterogeneous AuxSets, and further adopts contrastive learning to enforce loose coupling between datasets and high cohesion within single dataset to calibrate the TargetSet's representation space, thus mitigating knowledge confusion. Extensive experiments on five popular datasets across different domains demonstrate that DoKnowAD consistently outperforms existing TSAD baselines in various metrics.

1 Introduction

With the development of digital technologies (Chen et al. 2025; Li et al. 2025a; Wu et al. 2023, 2024b), real-time systems in domains like healthcare, industry, and finance are generating an increasing amount of time series data. These time series capture system dynamics and provide valuable insights into changes over time, enabling the detection of potential risks and threats. This approach, known as Time Series Anomaly Detection (TSAD), has gained growing attention from both academia and industry, with applications in health monitoring, manufacturing, and fraud detection, etc. (Schmidl, Wenig, and Papenbrock 2022).



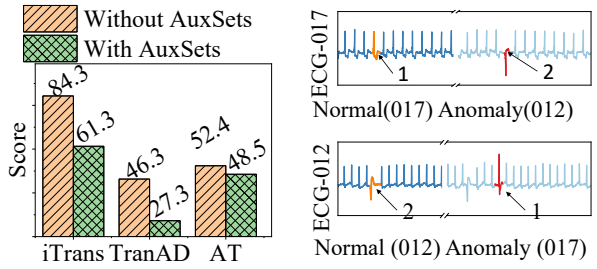
(a) Normal patterns are misclassified. (b) Similar Patterns.

Figure 1: Lack of domain knowledge leads to false alarms.

As time series data continue to grow and manual labeling remains costly, TSAD methods are increasingly moving toward unsupervised settings (Blázquez-García et al. 2021). Compared to traditional methods such as classification-based (Tax and Duin 2004) and density-estimation methods (Breunig et al. 2000; Tang et al. 2002), deep learning methods (Zamanzadeh Darban et al. 2024; Fu et al. 2025; Huang et al. 2025; Li et al. 2025b; Wu et al. 2022b, 2024a, 2025) offer stronger representation learning capabilities and have attracted more attention. These methods can be broadly categorized into two types: one reconstructs or predicts outputs using inputs and detects anomalies by comparing outputs with expected normal values, while the other measures the difference between test and normal samples in the representation space. In both cases, capturing normal behavior accurately is essential. To this end, various models based on GANs, VAEs, Transformers, GNNs, and Diffusions have been proposed (Li et al. 2019; Wang et al. 2024c,a; Yue et al. 2024; Zhang, Zhang, and Tsung 2022), aiming to model complex temporal dependencies embedded in time series, thereby constructing discriminative representations of normal patterns to improve detection performance.

However, deep learning-based TSAD methods often struggle to distinguish true anomalies from distribution shifts or pattern variations in test data, as they rely heavily

*Tao Ren is the Corresponding author.



(a) Results for ECG-017. (b) Anomaly Confusion.

Figure 2: Impact of directly incorporating AuxSets.

on training patterns that may not reflect domain definitions of anomalies. As shown in Figure 1 (a), the training set of ECG-805 from TSB-UAD (Paparrizos et al. 2022b) primarily consists of normal pattern-1. Methods like OmniAno (Xu et al. 2018) and AnoTrans (Xu et al. 2021) learn representations from pattern-1 and assign high anomaly scores to any deviation, as shown in the third row of Figure 1(a). While they correctly detect true anomalies, they also misclassify normal pattern-2 (*green shade*) due to representation shifts.

However, such fluctuations are generally normal, which could be seen as specific domain knowledge and should be flagged as normal. This observation highlights the necessity of integrating external domain knowledge to enhance the model's understanding of domain-normal patterns, rather than treating all test deviations from the training distribution as anomalies. In fact, similar patterns (*green lines*) also appear in other training datasets within the domain, as shown in Figure 1 (b), suggesting the potential of leveraging the auxiliary datasets (**AuxSets**) to assist the target dataset (**TargetSet**) to capture domain-level normality.

Whereas, when directly incorporating such domain knowledge from all available AuxSets to jointly construct a domain-level normal representation, we observe a decline in detection performance. As shown in Figure 2 (a), we conduct experiments using TimeMixer (Wang et al. 2024b), iTrans (Liu et al. 2024), TranAD (Tuli, Casale, and Jennings 2022), and AnoTrans (Xu et al. 2021), training them on ECG domain from UCR archive (Wu and Keogh 2021) and evaluating on ECG-017. Surprisingly, methods trained without AuxSets consistently outperform those incorporating them. To better understand this, we visualize datasets within the ECG domain, as shown in Figure 2 (b). Anomalies in ECG-017 resemble normal patterns in ECG-012, and vice versa, leading to semantic confusion of anomalies during joint representation. This misalignment between anomaly semantics in the TargetSet and AuxSets contributes to performance degradation. These findings suggest that directly incorporating all domain knowledge from AuxSets without identifying what is helpful may suppress the detection sensitivity to TargetSet-specific anomalies, highlighting a key challenge:

How to identify and incorporate helpful domain knowledge from AuxSets to enrich TargetSet's normal representation, while preserving its specific anomaly semantics.

To answer the question, we conduct an in-depth analysis on the ECG domain to investigate how domain knowledge from various AuxSets influences the TargetSet's representational ability and detection performance. As shown in Figure 3, by visualizing the normal and anomalous representations of the TargetSet after incorporating different AuxSets, we identify two phenomena: **First**, some AuxSets are inherently homogeneous with the TargetSet, reflecting similar perceptions of normal patterns. Under unsupervised learning, such limited domain-level diversity may adversely affect detection by reinforcing TargetSet-specific normality and making anomalies harder to distinguish. **Second**, while heterogeneous AuxSets broaden the TargetSet's understanding of domain-level normality, inconsistent anomaly semantics may cause TargetSet anomalies to drift into AuxSets' normal regions, weakening sensitivity to TargetSet-specific anomalies and degrading accurate detection.

Therefore, we formulate the *domain-oriented TSAD* problem and propose **Do**main **Know**ledge Enhanced Time Series **A**nomaly **D**etection (**DoKnowAD**), to address these issues through two core components. *First*, We propose an AuxSets refinement module that quantifies each AuxSet's impact based on the representation shift it induces in the TargetSet. Larger shifts suggest higher heterogeneity and can provide more diverse domain knowledge. *Second*, we design a representation calibration module that further incorporates the refined AuxSets to assist the TargetSet's representation learning. Specifically, we construct normality prototypes from the TargetSet's own representation rather than the joint representation, preserving its anomaly semantics while still benefiting from domain knowledge. The main contributions are summarized as follows:

- We propose leveraging AuxSets to enhance the TargetSet's normality representation. Through in-depth analysis, we find that homogeneous AuxSets provide limited knowledge, and furthermore heterogeneous AuxSets could make confusion for TargetSet detection.
- We propose DoKnowAD, which introduces a Representation HyperVolume Estimation metric to identify helpful heterogeneous AuxSets via representation space discrepancy, enhancing domain knowledge diversity.
- We design a contrastive loss to enforce loose TargetSet–AuxSets coupling and high TargetSet cohesion, enabling domain knowledge to calibrate the TargetSet's representation toward its own normality.
- DoKnowAD achieves up to 20% improvement over state-of-the-art methods on datasets from multiple domains, with further experiments highlighting its robustness across varied AuxSets and representation model.

2 Related Work

Time Series Anomaly Detection

Mainstream unsupervised TSAD methods can be broadly categorized by their detection space into two types: data-space and representation-space. Data-space methods (e.g., VAE (Su et al. 2019), GAN (Geiger et al. 2020), GNN (Zhang, Zhang, and Tsung 2022)) detect anomalies by comparing predicted or reconstructed and actual values.

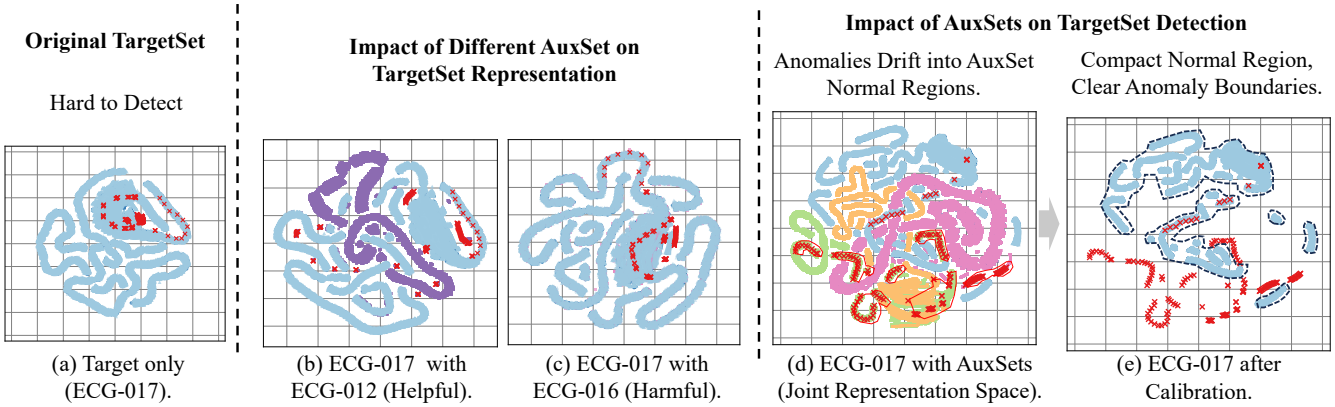


Figure 3: t-SNE visualization of the impact of different AuxSets on the representation and detection of TargetSet. (a) Original TargetSet. (b)-(c) Impact of different AuxSet on TargetSet Representation. (d)-(e) Impact of AuxSets on TargetSet Detection.

Representation-space methods identify anomalies by measuring discrepancies directly from learned representation, often using attention (Xu et al. 2021; Yang et al. 2023; Yue et al. 2024) or graph-based models (Deng and Hooi 2021). However, most of these methods are trained on individual datasets, resulting in incomplete normality representations. In contrast, our method incorporates helpful AuxSets to inject domain knowledge into the TargetSet, enhancing domain detection performance.

Time Series Representation

Time series representation has drawn increasing interest, particularly in forecasting tasks. Recent advances move beyond conventional black-box models (e.g., CNNs (Wang et al. 2023), RNNs (Lai et al. 2018), Transformers (Nie et al. 2023), and MLPs (Zeng et al. 2023)) by incorporating classical time series techniques like decomposition and multi-periodicity (Wang et al. 2024b), leading to improved downstream performance, including anomaly detection. While most models are trained on a single dataset, their architectures offer insights for modeling patterns. Building on this, DoKnowAD leverages AuxSets to enhance the TargetSet's representation and improve detection performance.

3 Motivation

Domain knowledge can help the TargetSet distinguish true anomalies rather than simple deviations from training patterns, but not all datasets in AuxSets are helpful and some maybe harmful. Using a dataset from ECG domain, We analyze how different AuxSets affect the TargetSet's representation and detection performance, respectively, as shown in Figure 3.

Impact of AuxSets on TargetSet Representation

As shown in Figure 3(a), training on ECG-017 alone results in most anomalies (*yellow circles*) near the normal region, increasing misclassification risk. Incorporating ECG-012, which differs from ECG-017, compresses the normal region and pushes anomalies (*green circles*) further away, improving separability. In contrast, adding ECG-016, which shares highly similar patterns with ECG-017, causes little change in the representation space. Due to the lack of helpful domain knowledge, anomaly separation becomes even poorer (*red Xs*). These results highlight the importance of refining AuxSets, as homogeneous AuxSets may degrade detection performance. Despite improvements from ECG-012, many anomalies from ECG-017 remain in the normal region of ECG-012 (*purple dots*), consistent with Figure 2.

Impact of AuxSets on TargetSet Detection

After incorporating heterogeneous AuxSets, we further investigate how this affects the detection performance. Taking ECG-017 as an example, we compare its test-time representation space before and after incorporating various AuxSets. As shown in Figure 3(a) and (d), the joint representation space after incorporating AuxSets shows that many anomalies in ECG-017 drift into the AuxSets normal regions (*red boundaries*), leading them to be normal at the domain level. This also explains why training without AuxSets outperforms training with them in Figure 2 (a). In contrast, removing the AuxSets' normals, as illustrated in Figure 3 (e), anomalies in ECG-017 results in better separation of anomalies from the normal cluster (*black dash boundaries*). This suggests that incorporating AuxSets to calibrate the TargetSet representation can effectively enlarge the boundary between normal and anomalous patterns.

Conclusion and Challenges

Through representation-level analysis, we identify **two key challenges** in leveraging AuxSets as domain knowledge. First, **homogeneous** AuxSets fail to introduce domain-level diversity, reinforcing TargetSet-specific normality and obscuring the boundary between normal and anomalous patterns. Second, heterogeneous AuxSets may cause TargetSet anomalies to **drift into the AuxSets' normal regions**, leading to semantic confusion and reduced detection accuracy.

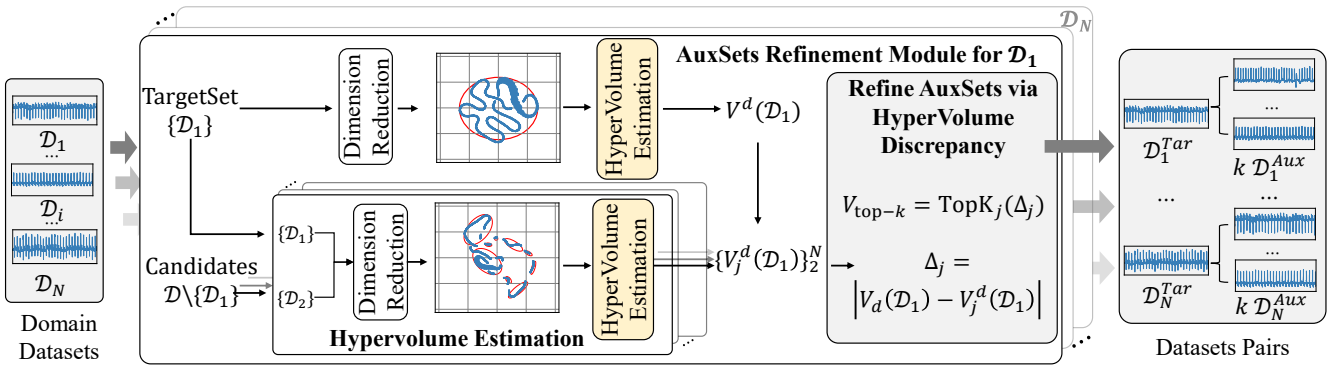


Figure 4: Overview of the AuxSets Refinement Module via Representation Hypervolume Discrepancy. After refinement, each TargetSet is paired with k heterogeneous AuxSets to incorporate the most helpful domain knowledge.

4 Method

To address these challenges, we reformulate TSAD into a domain-oriented problem and propose DoKnowAD, which tackles them from two complementary aspects. (1) It proposes an **AuxSets Refinement Module** through Representation HyperVolume Estimation to identify helpful heterogeneous AuxSets via representation space discrepancy. (2) It designs a **Representation Calibration Module** to enforce loose TargetSet–AuxSets coupling and high TargetSet cohesion, enabling domain knowledge to calibrate the TargetSet's representation toward its own normality.

Domain-Oriented TSAD Problem

Let $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$ be a collection of time series datasets from a specific domain, where each dataset $\mathcal{D}_n = \{X_1, X_2, \ldots, X_{L_n}\} \in \mathbb{R}^{L_n \times F}$ represents a time series with length L_n and each $X_{l_n} \in \mathbb{R}^F$ represents a fixed number of features F, which is consistent across all datasets in the domain. For simplicity, we drop the dataset-specific index n and use X_L to denote the time series data across all datasets. TSAD aims to design a model $\mathcal{F}$ that outputs anomaly scores $\mathbf{a} = \{a_1, a_2, \ldots, a_L\}$, where L is the total number of time steps across all datasets. A threshold δ_a is applied to predict the anomaly labels $\hat{Y} = \{y_1, y_2, \ldots, y_L\}$, where $y_l \triangleq 1_{\{a_l \geq \delta_a\}}$. To more fairly evaluate the performance of anomaly detection, we also assess the anomaly scores using threshold-independent metrics, such as ROC and VUS (Paparrizos et al. 2022a), etc .

AuxSets Refinement Module

As shown in Figure 4, to identify heterogeneous AuxSets for TargetSet, we first propose a **Representation HyperVolume Estimation** metric to quantify their impact by treating the representation shifts caused by AuxSets as **HyperVolume Discrepancy**. From Figure 3(b)–(d), we observe that heterogeneous AuxSets induce greater representation shifts and promote more compact structures. Based on this, AuxSets with larger hypervolume discrepancy are regarded as more heterogeneous and selected to provide diverse helpful domain knowledge, ultimately improving the TargetSet's normality modeling and anomaly separability.

Representational HyperVolume Estimation. Due to the high dimensionality of normal representations, direct geometric analysis is intractable. To address this, we project data into a lower-dimensional latent space using non-linear dimensionality reduction, which preserves structure and allow consistent comparison. Specifically, for each TargetSet $\mathcal{D}_i$ in the domain, we treat others $\{\mathcal{D}_j\}_{j \neq i}$ as candidate AuxSets. Each dataset is embedded via a mapping $\Phi(\cdot)$:

$$Z(\mathcal{D}_i) = \Phi(\mathcal{D}_i), \quad Z(\mathcal{D}_i \cup \mathcal{D}_j) = \Phi(\mathcal{D}_i \cup \mathcal{D}_j), \tag{1}$$

where $Z(\cdot) \subset \mathbb{R}^d$ denotes the low-dimensional representation. We fit a hyper-ellipsoid to the embedded TargetSet using its principal axes and compute its volume:

$$\mathrm{Vol}_d(Z) = \frac{\pi^{d/2}}{\Gamma(d/2+1)} \cdot \prod k = 1^d a_k, \tag{2}$$

where a_k is the semi-axis length along the k-th principal direction of the hyper-ellipsoid. The corresponding hypervolumes for the TargetSet alone and jointly with AuxSets are defined as:

$$\begin{aligned} V^d(\mathcal{D}_i) &= \mathrm{Vol}_d(Z(\mathcal{D}_i)), \\ V_j^d(\mathcal{D}_i) &= \mathrm{Vol}_d(Z(\mathcal{D}_i) \mid Z(\mathcal{D}_i, \mathcal{D}_j)), \end{aligned} \tag{3}$$

where the conditional notation reflects the joint embedding.

Refine Domain AuxSets via HyperVolume Discrepancy. Based on the above hypervolume metrics, we compute the absolute discrepancy for each candidate:

$$\Delta_j^{\mathrm{Vol}} = \left| V^d(\mathcal{D}_i) - V_j^d(\mathcal{D}_i) \right|, \tag{4}$$

which measures how much the inclusion of $\mathcal{D}j$ alters the spatial configuration of the target set. We rank all candidates $\{\mathcal{D}_j\}_{j \neq i}$ by their Δ_j^{Vol} and select the top-k with the largest discrepancies:

$$\mathcal{A}_{\text{top-}k}^{(i)} = \mathrm{TopK}_j \left(\Delta_j^{\mathrm{Vol}} \right). \tag{5}$$

This strategy prioritizes AuxSets that induce substantial structural shifts in the TargetSet's representation, improving coverage and generalization. As a result, the original dataset collection $\{\mathcal{D}_i\}_{i=1}^N$ is transformed into pairs $\{(\mathcal{D}_i^{\mathrm{Tar}}, \mathcal{D}_i^{\mathrm{Aux}})\}_{i=1}^N$, where each target set is paired with refined AuxSets by the proposed strategy.

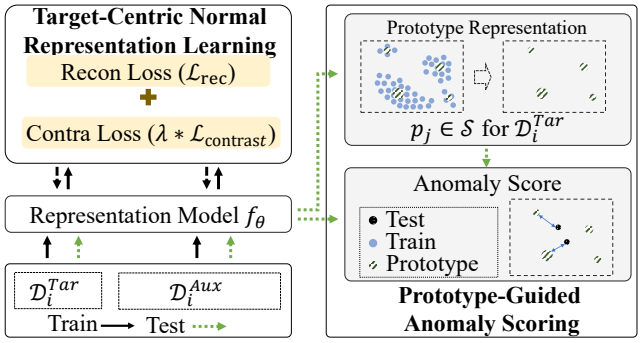


Figure 5: Representation Calibration and Scoring.

Representation Calibration and Scoring

Even after refinement, incorporating heterogeneous AuxSets can still introduce semantic confusion. As illustrated in Figure 3(d) and (e), anomalies in the TargetSet may drift into the normal regions defined by the AuxSets due to inconsistent anomaly semantics. This weakens the model's sensitivity to TargetSet-specific anomalies and hinders accurate detection.

To address this, DoKnowAD performs **Target-Centric Normal Representation Learning** on time series representation model during training and applies **Prototype-Guided Anomaly Scoring** during detection, as illustrated in Figure 5. When training, DoKnowAD treats TargetSet samples as positives and AuxSet samples as negatives in a contrastive learning setup, pulling TargetSet representations closer together while pushing them away from those of the AuxSets. This yields a more compact and discriminative representation space focused on the TargetSet. During detection, DoKnowAD constructs the TargetSet's normal prototype using refined representations, avoiding unrelated domain-level patterns and reducing misclassifications.

Target-Centric Normal Representation Learning.

In the representation model f_θ, input data x from the TargetSet and AuxSets is embedded into z and then reconstructed as $\hat{x}$. To enable calibration and ensure the model fits entire dataset pair, the overall objective combines reconstruction loss $\mathcal{L}_{\text{rec}}(x, \hat{x})$ and contrastive loss $\mathcal{L}_{\text{contrast}}(z_{\text{tar}}, z_{\text{aux}})$:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{rec}} + \lambda \cdot \mathcal{L}_{\text{contrast}}, \\ \mathcal{L}_{\text{contrast}} &= \mathcal{L}_{\text{tar}} - \mathcal{L}_{\text{tar-aux}}, \end{aligned} \tag{6}$$

where λ is a weighting hyperparameter, and $\mathcal{L}_{\text{tar}}$ and $\mathcal{L}_{\text{tar-aux}}$ denote contrastive losses that encourage high intra-TargetSet cohesion and loose TargetSet-AuxSets coupling, respectively. More details of f_θ and $\mathcal{L}$ are in Appendix.

Prototype-Guided Anomaly Scoring.

After training, we freeze the representation model to construct prototypes from the TargetSet's own representations, thereby preserving its dataset-specific normality. Specifically, we apply DBSCAN clustering to the embeddings z of all training instances, where each cluster center forms prototypes $p_j \in \mathcal{S}$. The total prototypes $\mathcal{S}$ is determined automatically by the clustering algorithm.

Domain	#Datasets	Avg Samples per Dataset		Avg Anomaly Ratio
		Train	Test	
ECG	22	10523	41095	0.70%
EPG	21	2786	24169	0.40%
SMD	38	5880	18327	2.63%
SMAP	19	1917	5783	3.74%
IOPS	17	5247	67545	1.59%

Table 1: Benchmark Details.

Given a TargetSet test data x^{test}, its anomaly score AS is computed as the distance to the nearest prototype:

$$AS(x^{\text{test}}) = \min_{p_j \in \mathcal{S}} \|f_\theta(x^{\text{test}}) - \mathbf{p}_j\|_2, \tag{7}$$

with higher scores indicating more likely anomalies.

5 Experiments

Benchmark Datasets

We select datasets from different domains, each characterized by complex modalities and widely adopted in literatures. These domains typically encompass a large number of datasets and align with the current development of benchmarks. Specifically, we include domains such as Health, Sensors, Environment, and Web Services. For instance, we use ECG and EPG from UCR (Wu and Keogh 2021) , as well as SMD, SMAP, and IOPS from TSB-AD (Liu and Paparrizos 2024). Detailed statistics of the datasets are in Table 1.

Baselines

We evaluate the performance of DoKnowAD by comparing it against 11 existing unsupervised methods. These baselines can be roughly divided into two groups. The *first group* focuses on detecting in data space. This includes models based on *basic architectures*, CNN(Munir et al. 2018) and LSTM (Hundman et al. 2018); *stochastic learning methods* including Donut (Xu et al. 2018) and OmniAnomaly (Su et al. 2019), *latent variable reconstruction models* like USAD (Audibert et al. 2020), *transformer-based models* including AnomalyTransformer (AnoTrans) (Xu et al. 2021), TranAD (Tuli, Casale, and Jennings 2022), and DCDetector (Yang et al. 2023), and *diffusion models* such as D3R (Wang et al. 2024a). The *second group* focuses on detecting in representation space, including M2N2 (Kim, Park, and Choo 2024) and DADA (Shentu et al. 2025).

Evaluation Metrics

The evaluation metrics for TSAD are generally divided into two categories. One line of metrics treats the task as a classification problem and relies on thresholding, such as F1-PA. However, the F1-PA metric tends to overestimate classifier performance (Kim et al. 2022), even though it has certain practical ustifications. In this work, we adopt the affiliation F1 score (F1-A) as one of the evaluation metrics. On the other hand, to assess the discriminative power of anomaly scores, additional threshold-independent metrics are also used, including range-based AUC-PR (**R_A_P**), VUS-ROC

Method	ECG		EPG		SMD		SMAP		IOPS	
	F1-A	V_R	F1-A	V_R	F1-A	V_R	F1-A	V_R	F1-A	V_R
CNN [2018, IEEE]	73.49	71.18	75.55	64.57	95.05	**90.19**	91.06	77.24	75.97	82.34
LSTM [2018, KDD]	72.56	59.94	72.58	66.94	94.22	85.83	92.04	65.27	73.27	80.65
USAD [2020, KDD]	75.73	57.55	79.87	63.36	84.36	87.84	94.52	71.21	56.42	77.14
Donut [2018, WWW]	69.18	55.42	73.54	64.68	85.58	71.77	84.57	73.44	67.96	71.88
OmniAnomaly [2019, KDD]	74.51	53.41	79.77	62.06	79.18	86.40	85.43	79.25	53.68	74.96
AnoTrans [2021, ICLR]	68.60	52.11	69.41	51.54	78.35	79.72	95.08	51.03	55.64	75.17
TranAD [2022, VLDB]	70.98	55.07	73.28	65.96	86.98	79.45	89.64	59.88	62.15	64.29
DCDetector [2023, KDD]	67.13	53.78	72.19	64.97	85.29	81.20	94.26	66.31	59.09	71.09
M2N2 [2024, AAAI]	73.60	63.00	69.09	65.63	91.86	86.67	95.82	78.62	71.10	84.98
D3R [2024a, NIPS]	65.77	59.05	64.92	51.29	89.12	78.01	88.76	76.81	72.83	69.93
DADA [2025, ICLR]	67.23	54.67	65.21	57.93	94.35	84.10	95.69	73.76	79.47	72.85
DoKnowAD (Ours)	**90.15**	**75.30**	**82.46**	**68.53**	**95.41**	87.72	**97.24**	**79.63**	**83.41**	**87.74**

Table 2: Performance comparison of baselines. All results are averaged over 5 independent trials for each metric. All results are in %. The best results are highlighted in bold, and the second ones are underlined.

Var.	Components				ECG		EPG		SMD		SMAP		IOPS	
	AuxSets	Refine	Contra	Proto	F1-A	V_R	F1-A	V_R	F1-A	V_R	F1-A	V_R	F1-A	V_R
1	✗	✗	✗	✗	56.12	57.52	64.24	61.20	93.22	86.82	94.13	73.64	79.64	82.25
2	✓	✗	✓	✓	64.49	60.82	66.07	61.28	94.16	86.47	94.87	74.07	82.67	85.93
3	✓	✓	✗	✓	82.49	72.71	75.68	64.18	94.62	86.73	96.94	76.68	83.12	86.89
4	✓	✓	✓	✗	69.49	57.82	79.66	58.06	93.21	87.19	95.66	76.52	82.54	85.10
5	✓	✓	✓	✓	**90.15**	**75.30**	**82.46**	**68.53**	**95.41**	**87.72**	**97.24**	**79.63**	**83.41**	**87.74**

Table 3: Ablation studies on components of DoKnowAD, including AuxSets, refinement strategy (Refine), contrastive learning (Contra), and prototype-based detection (Proto). All results are reported in %, with the best highlighted in bold.

(**V_R**), and VUS-PR (**V_P**) (Paparrizos et al. 2022a). In the main paper, we report F1-A and V_R as the primary metrics for SOTA comparison.

Main Results

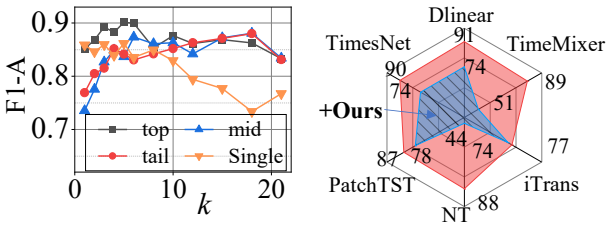
Each domain includes multiple datasets, as detailed in Table 1. For all baselines, we perform training and detection independently for each domain. Both threshold-dependent (F1-A) and threshold-independent (V_R) evaluation metrics are used. To ensure fairness and avoid manual tuning of hyperparameters such as SPOT (Siffer et al. 2017), we apply grid search to determine the optimal threshold for each method. The overall results are shown in Table 2. DoKnowAD consistently achieves the best performance under both evaluation metrics. In domains with complex patterns, such as ECG and EPG in the health category, it surpasses other methods by over 20%. For more stable domains like SMD, SMAP, and IOPS, which are widely used in prior work, the performance gap between methods is smaller, suggesting these datasets are less challenging.

To evaluate the effect of domain knowledge on baselines, *we further equip each baseline with All and Refined AuxSets, respectively*, and results are provided in Appendix D. Although most baselines still fall behind DoKnowAD, they show some improvements when augmented with AuxSets. However, on the SMD dataset, including All AuxSets leads to performance degradation for most baselines, with performance drops of up to 3%. This is because the too much AuxSets in SMD causes the model to lose sensitivity to dataset-specific anomalies. These findings confirm that adding helpful AuxSets is crucial for performance.

Ablation study

We further *evaluate the contribution of each component* in Table 3. We first consider the original representation model, where each TargetSet is trained independently (Variant 1). This variant performs poorly on datasets such as ECG, indicating that limited training data without domain knowledge leads to sub-optimal performance. On the other hand, incorporating all AuxSets without refinement (Variant 2) still slightly improves performance, confirming the effectiveness of domain knowledge. However, excessive domain knowledge may over-compress the TargetSet's representation, leading to dense prototypes and reduced anomaly separability, underscoring the need for refinement. Notably, when AuxSets are refined, using them to assist prototype construction (Variant 3) yields better performance than contrastive training alone (Variant 4), highlighting the importance of prototype construction. Relying solely on contrastive learning during training, without considering the detection phase, may weaken the model's ability to understand dataset-specific normal patterns.

To validate the effectiveness of our refinement module, we *evaluate multiple time series representation models* under different configurations, including AuxSets, Refine, and Calibrate variants. As provided in Appendix D, although



(a) Refinement Strategy. (b) Representation Models.

Figure 6: (a) Analysis of the AuxSets Refinement Strategy and (b) Impact of DoKnowAD on different representation models for ECG domain datasets.

performance varies across models and domains due to architectural differences, most benefit considerably from the incorporation of domain knowledge.

Method Analysis

To further evaluate DoKnowAD in terms of knowledge refinement and representation calibration, we conduct analyses from two perspectives: selecting AuxSets with different strategies, and applying different representation models, as shown in Figure 6.

Analysis on the AuxSets Refinement Strategy We rank candidate datasets for each TargetSet using hypervolume metrics and select the top-k as AuxSets. To validate the refinement strategy, *we compare three selection variants: top-k, mid-k, and tail-k*, selecting the k datasets with the largest, middle, and smallest hypervolume discrepancies, respectively. We also evaluate the contribution of each individual AuxSet (*single*), as shown in Figure 6(a). Results show that top-k performs best, but performance declines beyond a certain k, suggesting that excessive AuxSets may over-compress the TargetSet's representation. Nonetheless, it still outperforms the without-AuxSet baseline (56.12), while other variants perform worse, demonstrating the effectiveness of our refinement strategy.

Improvements on other Representation Models On the other hand, *we evaluate DoKnowAD across a range of state-of-the-art representation models*, including DLinear (Zeng et al. 2023), iTransformer (Liu et al. 2024), TimeMixer (Wang et al. 2024b), PatchTST (Nie et al. 2023), Nonstationary Transformer (Liu et al. 2022), and TimesNet (Wu et al. 2022a). By integrating DoKnowAD into these backbones, we aim to assess its generalizability and its ability to enhance anomaly detection across different model architectures. As shown in Figure 6 (b), DoKnowAD consistently improves detection performance across all models (up to 40%+). More detailed results and discussions for other domain datasets are provided in Appendix.

Visualization Analysis

To intuitively analyze DoKnowAD, *we visualize anomaly scores on three representative domains*: ECG, EPG, and SMD. We compare three variants: (1) Only-TargetSet, using reconstruction-based detection without any AuxSets; (2) DoKnowAD w/o Refine, using all domain AuxSets without refinement and also uses reconstruction scores; and (3) DoKnowAD, which using AuxSets refinement module and representation calibration module.

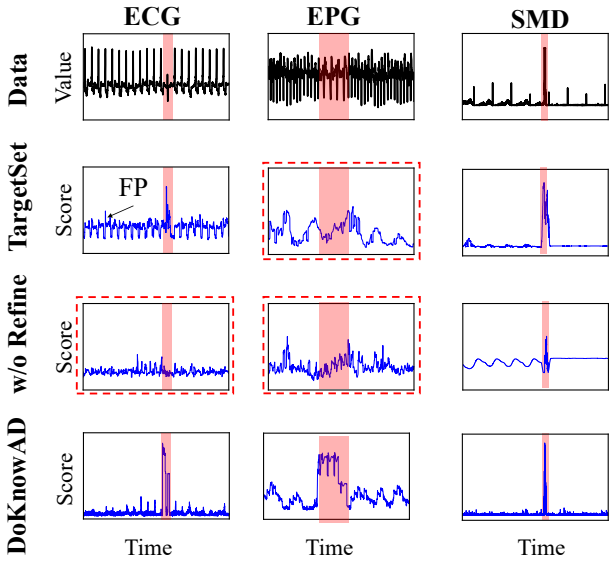


Figure 7: Anomaly score visualizations across different datasets. Failure cases are highlighted with red triangles.

As shown in Figure 7, for complex domains such as ECG and EPG, the Only-TargetSet approach fails to yield clearly separable scores, and directly introducing all AuxSets (w/o Refine) may even degrade performance due to the decline of discriminability in anomalous segments. In contrast, DoKnowAD achieves high separation by effectively incorporating domain knowledge while preserving dataset-specific anomaly patterns. For relatively simpler domains like SMD, where anomalies tend to be large outliers, the performance gap among variants is small. Nonetheless, DoKnowAD still yields superior anomaly distinction, validating its general effectiveness across diverse domain scenarios.

6 Conclusion

In this paper, we proposed DoKnowAD, a framework for time series anomaly detection that incorporates refined domain knowledge from AuxSets to enhance the representations of normal patterns while maintaining sensitivity to dataset-specific anomalies in the TargetSet. DoKnowAD significantly improves detection performance, especially in domains with complex patterns such as ECG and EPG. Extensive experiments across multiple domains and representation models show that DoKnowAD consistently outperforms SOTA baselines under both threshold-dependent and independent metrics. In the future, we aim to build intra-domain and inter-domain representation models to achieve more generalizable time series anomaly detection, instead of refining AuxSets separately for each TargetSet, thereby enhancing generalization and reducing computational cost.

Acknowledgements

This work was supported by the Zhejiang Province Key R&D Program of China (Grant No. 2024C01071), National Natural Science Foundation of China (Grant No. U23B2025), and the National Science and Technology Major Project of China (Grant No. 2024ZD1401806).

References

Audibert, J.; Michiardi, P.; Guyard, F.; Marti, S.; and Zuluaga, M. A. 2020. USAD: Unsupervised anomaly detection on multivariate time series. In *ACM SIGKDD*, 3395–3404.

Blázquez-García, A.; Conde, A.; Mori, U.; and Lozano, J. A. 2021. A review on outlier/anomaly detection in time series data. *CSUR*, 54(3): 1–33.

Breunig, M. M.; Kriegel, H.-P.; Ng, R. T.; and Sander, J. 2000. LOF: identifying density-based local outliers. In *ACM SIGMOD*, 93–104.

Chen, Z.; Hu, Y.; Li, Z.; Fu, Z.; Song, X.; and Nie, L. 2025. OFFSET: Segmentation-based focus shift revision for composed image retrieval. In *ACM MM*, 6113–6122.

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

Fu, Z.; Li, Z.; Chen, Z.; Wang, C.; Song, X.; Hu, Y.; and Nie, L. 2025. PAIR: Complementarity-guided Disentanglement for Composed Image Retrieval. In *ICASSP*, 1–5. IEEE.

Geiger, A.; Liu, D.; Alnegheimish, S.; Cuesta-Infante, A.; and Veeramachaneni, K. 2020. Tadgan: Time series anomaly detection using generative adversarial networks. In *IEEE Big Data*, 33–43. IEEE.

Huang, Q.; Chen, Z.; Li, Z.; Wang, C.; Song, X.; Hu, Y.; and Nie, L. 2025. MEDIAN: Adaptive Intermediate-grained Aggregation Network for Composed Image Retrieval. In *ICASSP*, 1–5. IEEE.

Hundman, K.; Constantinou, V.; Laporte, C.; Colwell, I.; and Soderstrom, T. 2018. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *ACM SIGKDD*, 387–395.

Kim, D.; Park, S.; and Choo, J. 2024. When model meets new normals: test-time adaptation for unsupervised time-series anomaly detection. In *AAAI*, volume 38, 13113–13121.

Kim, S.; Choi, K.; Choi, H.-S.; Lee, B.; and Yoon, S. 2022. Towards a rigorous evaluation of time-series anomaly detection. In *AAAI*, volume 36, 7194–7201.

Lai, G.; Chang, W.-C.; Yang, Y.; and Liu, H. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In *ACM SIGIR conference on research & development in information retrieval*, 95–104.

Li, D.; Chen, D.; Jin, B.; Shi, L.; Goh, J.; and Ng, S.-K. 2019. MAD-GAN: Multivariate anomaly detection for time series data with generative adversarial networks. In *ICANN*, 703–716. Springer.

Li, Z.; Chen, Z.; Wen, H.; Fu, Z.; Hu, Y.; and Guan, W. 2025a. ENCODER: Entity mining and modification relation binding for composed image retrieval. In *AAAI*, volume 39, 5101–5109.

Li, Z.; Fu, Z.; Hu, Y.; Chen, Z.; Wen, H.; and Nie, L. 2025b. FineCIR: Explicit parsing of fine-grained modification semantics for composed image retrieval. *arXiv:2503.21309*.

Liu, Q.; and Paparrizos, J. 2024. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. *NIPS*, 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. *ICLR*.

Liu, Y.; Wu, H.; Wang, J.; and Long, M. 2022. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *NIPS*.

Munir, M.; Siddiqui, S. A.; Dengel, A.; and Ahmed, S. 2018. DeepAnT: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, 7: 1991–2005.

Nie, Y.; Nguyen, N. H.; Sinthong, P.; and Kalagnanam, J. 2023. A time series is worth 64 words: Long-term forecasting with transformers. *ICLR*.

Paparrizos, J.; Boniol, P.; Palpanas, T.; Tsay, R. S.; Elmore, A.; and Franklin, M. J. 2022a. Volume Under the Surface: A New Accuracy Evaluation Measure for Time-Series Anomaly Detection. *VLDB*, 15(11): 2774–2787.

Paparrizos, J.; Kang, Y.; Boniol, P.; Tsay, R. S.; Palpanas, T.; and Franklin, M. J. 2022b. TSB-UAD: an end-to-end benchmark suite for univariate time-series anomaly detection. *VLDB*, 15(8): 1697–1711.

Schmidl, S.; Wenig, P.; and Papenbrock, T. 2022. Anomaly detection in time series: a comprehensive evaluation. *VLDB*, 15(9): 1779–1797.

Shentu, Q.; Li, B.; Zhao, K.; Shu, Y.; Rao, Z.; Pan, L.; Yang, B.; and Guo, C. 2025. Towards a General Time Series Anomaly Detector with Adaptive Bottlenecks and Dual Adversarial Decoders. *ICLR*.

Siffer, A.; Fouque, P.-A.; Termier, A.; and Largouet, C. 2017. Anomaly detection in streams with extreme value theory. In *ACM SIGKDD*, 1067–1075.

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 *ACM SIGKDD*, 2828–2837.

Tang, J.; Chen, Z.; Fu, A. W.-C.; and Cheung, D. W. 2002. Enhancing effectiveness of outlier detections for low density patterns. In *PAKDD*, 535–548. Springer.

Tax, D. M.; and Duin, R. P. 2004. Support vector data description. 54: 45–66.

Tuli, S.; Casale, G.; and Jennings, N. R. 2022. Tranad: Deep transformer networks for anomaly detection in multivariate time series data. *VLDB*.

Wang, C.; Zhuang, Z.; Qi, Q.; Wang, J.; Wang, X.; Sun, H.; and Liao, J. 2024a. Drift doesn't Matter: Dynamic Decomposition with Diffusion Reconstruction for Unstable Multivariate Time Series Anomaly Detection. 36.

Wang, H.; Peng, J.; Huang, F.; Wang, J.; Chen, J.; and Xiao, Y. 2023. MICN: Multi-scale Local and Global Context Modeling for Long-term Series Forecasting.

Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J. Y.; and Zhou, J. 2024b. Timemixer: Decomposable multiscale mixing for time series forecasting. *ICLR*.

Wang, Z.; Pei, C.; Ma, M.; Wang, X.; Li, Z.; Pei, D.; Rajmohan, S.; Zhang, D.; Lin, Q.; Zhang, H.; et al. 2024c. Revisiting vae for unsupervised time series anomaly detection: A frequency perspective. In *ACM Web Conference 2024*, 3096–3105.

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

Wu, R.; and Keogh, E. J. 2021. Current time series anomaly detection benchmarks are flawed and are creating the illusion of progress. *TKDE*, 35(3): 2421–2429.

Wu, X.; Liu, X.; Niu, J.; Wang, H.; Tang, S.; Zhu, G.; and Su, H. 2024a. Decoupling general and personalized knowledge in federated learning via additive and low-rank decomposition. In *ACM MM*, 7172–7181.

Wu, X.; Liu, X.; Niu, J.; Zhu, G.; and Tang, S. 2023. Bold but cautious: Unlocking the potential of personalized federated learning through cautiously aggressive collaboration. In *CVPR*, 19375–19384.

Wu, X.; Niu, J.; Liu, X.; Ren, T.; Huang, Z.; and Li, Z. 2022b. pFedgf: Enabling personalized federated learning via gradient fusion. In *IPDPS*, 639–649. IEEE.

Wu, X.; Niu, J.; Liu, X.; Shi, M.; Zhu, G.; and Tang, S. 2024b. Tackling Feature-Classifier Mismatch in Federated Learning via Prompt-Driven Feature Transformation. *arXiv:2407.16139*.

Wu, X.; Niu, J.; Liu, X.; Zhu, G.; Tang, S.; Lin, W.; and Cao, J. 2025. The diversity bonus: Learning from dissimilar clients in personalized federated learning. *TNNLS*.

Xu, H.; Chen, W.; Zhao, N.; Li, Z.; Bu, J.; Li, Z.; Liu, Y.; Zhao, Y.; Pei, D.; Feng, Y.; et al. 2018. Unsupervised anomaly detection via variational auto-encoder for seasonal kpis in web applications. In *WWW*, 187–196.

Xu, J.; Wu, H.; Wang, J.; and Long, M. 2021. Anomaly transformer: Time series anomaly detection with association discrepancy.

Yang, Y.; Zhang, C.; Zhou, T.; Wen, Q.; and Sun, L. 2023. Dcdetector: Dual attention contrastive representation learning for time series anomaly detection. In *ACM SIGKDD*, 3033–3045.

Yue, W.; Ying, X.; Guo, R.; Chen, D.; Shi, J.; Xing, B.; Zhu, Y.; and Chen, T. 2024. Sub-Adjacent Transformer: Improving Time Series Anomaly Detection with Reconstruction Error from Sub-Adjacent Neighborhoods.

Zamanzadeh Darban, Z.; Webb, G. I.; Pan, S.; Aggarwal, C.; and Salehi, M. 2024. Deep Learning for Time Series Anomaly Detection: A Survey. *CSUR*, 57(1).

Zeng, A.; Chen, M.; Zhang, L.; and Xu, Q. 2023. Are transformers effective for time series forecasting? In *AAAI*, volume 37, 11121–11128.

Zhang, W.; Zhang, C.; and Tsung, F. 2022. Grelen: Multivariate time series anomaly detection from the perspective of graph relational learning. In *IJCAI*, 2390–2397.