EAI Endorsed Transactions
on Scalable Information Systems
Research Article EAI.EU

Privacy-Preserving SACAA-Net: A Scale-Aware Consistency and Anomaly-Attentive Network for Fine-Grained Industrial Anomaly Detection in Secure Distributed Edge-IIoT Systems

Jiani Zeng[1,*], Xinyu He[1], Weixian Wang[1], Chenyu Jin[1], Hui Xu[1], Siyue Lu[1]

[1]Institute of Artificial Intelligence, State Grid Beijing Electric Power Company, Beijing, 100031, China

Abstract

Fine-grained industrial anomaly detection has become a core component of intelligent inspection in Industrial Internet of Things (IIoT) scenarios, particularly when visual data are continuously acquired and processed through distributed sensors together with edge devices. Yet once industrial visual data and model updates are distributed in this way, a series of security risks becomes difficult to avoid, including data leakage, unauthorized access, gradient inversion attacks, model poisoning, and privacy inference. What may be exposed through these risks is not only raw data itself, but also sensitive information related to equipment structure, production status, and enterprise operations. Under such conditions, privacy preservation is no longer optional; it is an essential requirement for real-world industrial deployment. Existing methods still concentrate mainly on representation learning under heterogeneous data distributions, while security constraints in distributed environments are often insufficiently considered. This can easily result in inconsistent feature representations and inadequate protection of privacy-sensitive information during both training and transmission. To address these issues, we propose Privacy-Preserving SACAA-Net for fine-grained industrial anomaly detection in secure distributed IIoT systems. The proposed model brings together cross-scale semantic consistency learning, anomaly-aware attention, and structural dependency modeling, with the goal of improving detection accuracy while retaining fine-grained details and overall structural information. For privacy-preserving deployment, SACAA-Net supports local training on edge devices, reduces the need for raw data sharing, and remains compatible with secure aggregation, differential privacy, and encrypted transmission, thereby helping mitigate the risks of information leakage and model inversion attacks.Extensive experiments on six industrial datasets demonstrate that SACAA-Net achieves superior performance in both accuracy and robustness, achieving average image-level and pixel-level AUROC scores of 96.4% and 92.3%, respectively, and outperforming the strongest baseline by 1.5% and 1.8% on these two metrics, providing an effective and secure solution for distributed industrial anomaly detection.

Received on 06 May 2026; accepted on 01 August 2026; published on 24 August 2026

Keywords: privacy-preserving edge intelligence, secure industrial anomaly detection, distributed industrial IoT, fine-grained anomaly detection, data security

doi:10.4108/eetsis.12925

1. Introduction

In real-world distributed Industrial Internet of Things (IIoT) environments and edge-based industrial inspection systems[1, 2], fine-grained anomaly detection has become increasingly important for ensuring industrial reliability, operational safety, and secure system intelligence under privacy-sensitive and security-critical conditions. Unlike conventional centralized inspection paradigms, modern industrial monitoring systems

*Corresponding author: Jiani Zeng. Email: dkyai2026@126.com

continuously collect visual data through geographically distributed sensors, edge devices, and multi-camera platforms. Such distributed sensing infrastructures not only improve monitoring coverage and response efficiency but also introduce new concerns regarding secure data acquisition, privacy preservation, and adversarial robustness in collaborative industrial intelligence systems. Consequently, industrial anomaly detection is evolving from simple visual defect recognition toward privacy-aware, security-critical, and distributed intelligent inspection systems capable of maintaining robust detection performance under decentralized industrial environments.

In real-world industrial visual inspection scenarios[1, 2], anomalies do not usually present themselves as large and obvious global defects. More often, they emerge as localized irregularities at the fine-grained component level, such as micro-cracks, edge missing, structural deformation, or subtle disturbances in functional regions[3]. A notable property of these anomalies is their "weak signal" nature. In many cases, they occupy only a very limited spatial area within the image, while their visual variations do not produce sufficiently strong contrast. Because of this, they are easily confused with background textures or even normal structural patterns.Consequently, compared with traditional object detection or classification tasks, privacy-preserving and security-aware fine-grained anomaly detection in distributed industrial environments relies more heavily on precise local detail modeling and coherent global structural understanding.

From a system-level perspective, distributed industrial anomaly detection systems face fundamental security and privacy bottlenecks across the entire data lifecycle, including data acquisition, transmission, model training, and inference. In particular, industrial inspection scenarios such as State Grid infrastructure monitoring involve highly sensitive and security-critical data, including power equipment images, hidden danger records from distribution stations, patrol inspection logs, and defect annotations. These data directly reflect the operational state of critical infrastructure and therefore cannot be exposed to centralized cloud servers. Consequently, cross-domain data flows among heterogeneous edge devices and cloud servers significantly increase the risk of industrial information leakage, while resource-constrained edge nodes often lack sufficient capability to support strong encryption or secure computation protocols. Centralized training paradigms make privacy leakage risks even harder to control, since raw visual data may directly expose equipment structures, production configurations, and operational states. There is also a further concern in distributed industrial AI systems: once edge nodes are compromised, malicious gradients or poisoned updates may be injected into the training process, eventually causing model contamination and reduced detection reliability.

In practical distributed IIoT systems, industrial visual data are usually generated and processed at edge nodes close to physical devices, rather than being fully transmitted to centralized servers. This decentralized processing pattern does improve real-time responsiveness and helps reduce communication latency, but it also brings a new set of challenges involving security, privacy, and adversarial robustness during data collection, transmission, training, and inference. Sensitive industrial information may be exposed through centralized aggregation, communication leakage, or heterogeneous interactions among edge nodes, which makes privacy-preserving and secure anomaly analysis increasingly necessary for industrial edge intelligence systems. Under these conditions, anomaly detection models are expected to do more than perform accurate defect recognition; they also need to support robust learning, secure inference, and privacy-preserving collaboration in distributed, heterogeneous, and resource-constrained environments.

In modern industrial systems, visual data are often acquired by distributed sensors, edge devices, and multi-camera platforms deployed across different locations. Once data are collected in this way, a number of additional challenges emerge, including heterogeneous data distributions, limited communication resources, and inconsistent feature representations across devices. In practice, such heterogeneity may arise from many sources, such as different camera viewpoints, varying image resolutions, inconsistent illumination conditions, sensor-specific noise patterns, factory-dependent environmental dynamics, and diverse sampling frequencies across industrial edge nodes. Differences in hardware among edge devices, together with variations in deployment environments, frequently cause severe domain shifts and non-IID data distributions, which make feature alignment and collaborative anomaly modeling substantially more difficult in distributed industrial systems. Similar difficulties are further intensified by differences in sensing conditions, hardware configurations, and local environmental dynamics, as these factors often produce severe cross-device distribution shifts and complicate reliable anomaly modeling across decentralized industrial networks. As a result, anomaly detection models must not only capture fine-grained visual patterns, but also remain robust and consistent under distributed and resource-constrained environments.

Another important challenge lies in the limitations of existing industrial anomaly detection benchmarks.Although datasets such as MVTec AD and VisA have significantly promoted the development of industrial anomaly detection research, many current benchmarks still suffer from limited data scale, relatively

simple anomaly categories, insufficient environmental diversity, and weak support for cross-device or cross-factory generalization. In particular, most existing datasets are collected under relatively controlled laboratory conditions, which cannot fully reflect the heterogeneous sensing conditions and distributed deployment characteristics commonly encountered in real-world industrial IoT systems. Consequently, models trained on such benchmarks may exhibit degraded robustness when deployed in practical distributed industrial environments.

In this work, SACAA-Net is formulated under a privacy-preserving unsupervised industrial anomaly detection paradigm, where the model is trained primarily using normal samples without requiring large-scale pixel-level anomaly annotations during optimization. This design is particularly well suited to distributed industrial environments, where abnormal samples are usually scarce, difficult to annotate, and, in many cases, sensitive to share across edge nodes.

Existing studies often approach this problem through multi-scale feature learning[4], in which information from different resolutions is fused to strengthen anomaly perception. Although such strategies can, to some extent, compensate for the limitations of single-scale representations, they usually rely on an implicit assumption that features from different scales can be directly aligned and aggregated. In practice, this assumption is not always easy to satisfy. Low-level features are more likely to preserve local texture details, whereas high-level features tend to encode semantic structures, and the gap between these two representation spaces can be quite substantial. If no appropriate constraint is imposed, cross-scale fusion may bring semantic inconsistency rather than effective complementarity. Under such circumstances, local anomalies may be overwhelmed by dominant high-level semantics, while global structural information may also be weakened, ultimately reducing final detection performance. The problem becomes more pronounced in distributed industrial environments, where heterogeneous edge nodes may produce inconsistent semantic representations because acquisition conditions vary and data distributions remain decentralized.

Attention mechanisms have likewise been widely introduced to improve the model's ability to focus on informative regions[5–7]. Even so, many existing attention designs are still primarily driven by saliency or response intensity, which means they naturally favor regions that are easier to recognize. Such a strategy can work well for tasks like object detection, but it is less appropriate for anomaly detection.In many cases, anomalies correspond to weak but critical signals whose responses are not prominent[8, 9], especially in early feature representations where they may be indistinguishable from background noise. As a result, uniform attention allocation tends to overlook these regions, leading to missed detections or inaccurate localization. Different from conventional saliency-driven attention mechanisms that mainly emphasize dominant semantic regions with strong responses, anomaly-aware attention explicitly enhances low-response yet anomaly-sensitive regions by leveraging anomaly likelihood estimation. Such a design is motivated by the observation that many industrial defects exhibit weak local responses rather than globally salient patterns. Therefore, anomaly-aware modulation allows subtle abnormal cues to receive higher adaptive amplification during feature propagation, thereby improving the detectability of weak anomalies under complex industrial backgrounds. Furthermore, in privacy-sensitive and security-critical industrial settings, inaccurate anomaly localization may increase inference uncertainty across distributed nodes, thereby affecting the reliability of secure collaborative industrial monitoring systems.

To more clearly distinguish SACAA-Net from representative industrial anomaly detection methods such as PatchCore, PaDiM, and DRAEM, we summarize its main contributions from three complementary perspectives. First, the proposed cross-scale consistency learning explicitly addresses the semantic misalignment introduced by direct multi-scale fusion, especially under heterogeneous and non-IID edge environments, thereby reducing the risk that weak local anomalies are suppressed by dominant high-level semantics. Second, unlike conventional saliency-driven attention mechanisms[5–7] that mainly emphasize highly responsive or visually dominant regions, the proposed anomaly-aware attention is designed to selectively amplify low-response yet anomaly-sensitive cues, which is particularly important for fine-grained industrial defects with weak visual contrast[8, 9]. Third, structural dependency modeling is introduced not merely as an auxiliary refinement module, but as a necessary mechanism for preserving component-level relational coherence, since many industrial anomalies manifest as localized structural inconsistencies rather than isolated texture deviations.

Motivated by the above observations, this paper proposes Privacy-Preserving SACAA-Net, a unified distributed industrial anomaly detection framework that jointly integrates cross-scale semantic consistency learning, anomaly-aware attention enhancement, and structural dependency modeling under privacy-preserving distributed optimization constraints. Different from existing methods that independently optimize feature fusion, attention allocation, or structural reasoning, SACAA-Net emphasizes their synergistic optimization within a shared semantic representation space. Specifically, cross-scale consistency learning stabilizes heterogeneous feature representations

across distributed edge nodes, anomaly-aware attention selectively amplifies weak but critical defect signals, and structural dependency modeling preserves global relational coherence among industrial components. Through this collaborative optimization mechanism, SACAA-Net improves anomaly localization accuracy, semantic robustness, and deployment reliability at the same time under heterogeneous industrial IoT environments.

2. Related Work

2.1. Distributed Industrial Anomaly Detection

In distributed Industrial Internet of Things (IIoT) environments, privacy-preserving fine-grained anomaly detection places a dual demand on the model: it must be able to capture subtle local abnormalities while also maintaining semantic consistency across heterogeneous edge nodes. Under this setting, multi-scale feature learning has gradually become a core strategy for fine-grained anomaly detection, since the task depends on both local detail perception and global semantic understanding. Common approaches typically rely on feature pyramid structures or cross-layer fusion mechanisms to integrate representations from different resolutions. For example, feature pyramid networks (FPN) [10] pass high-level semantic information to lower layers through a top-down pathway, thereby strengthening the representation of small-scale patterns. Some methods also adopt multi-branch architectures, in which features at different scales are modeled separately before being aggregated for improved anomaly representation.

In anomaly detection tasks, multi-scale modeling is often used together with patch-based or reconstruction-based strategies. Patch-level methods[11] divide images into local regions and then perform similarity matching across multiple scales to localize anomalies. Reconstruction-based approaches[12, 13], on the other hand, exploit multi-scale reconstruction errors or feature distribution discrepancies to identify abnormal regions.

With the continued deployment of distributed sensing infrastructures in industrial IoT systems, recent research has paid increasing attention to distributed anomaly detection under edge-based settings, where visual data are generated and processed across geographically separated devices. The general purpose of these methods is to improve detection robustness, reduce communication overhead, and at the same time make local inference feasible at edge nodes.

Even with these advances, many existing methods still rest on an implicit assumption that features from different scales can be aligned and fused directly. In practice, this assumption is often difficult to satisfy. Low-level features usually preserve texture details, whereas high-level features tend to represent more abstract semantics. Once the task is placed in distributed settings, heterogeneous sensing conditions together with domain shifts across edge nodes further intensify feature inconsistency.

2.2. Privacy-Preserving Edge Intelligence and Secure Federated Learning

Attention mechanisms have been widely introduced to improve a model's ability to focus on informative regions through feature-response reweighting. Representative designs include channel attention, spatial attention, and hybrid forms that combine both. For instance, SENet improves feature representation by recalibrating channel-wise responses[14], while CBAM sequentially models[6] channel attention and spatial attention.

In distributed industrial environments, attention mechanisms have also been extended to privacy-aware edge intelligence settings, where feature extraction and inference are carried out locally in order to avoid direct exposure of raw data. Under this design, edge nodes exchange compact feature representations rather than raw images, which makes collaborative perception possible while still preserving privacy.

Beyond edge intelligence itself, secure federated learning and distributed optimization have also been extensively studied for privacy-preserving collaborative training. These methods generally depend on local model training, encrypted communication, and secure aggregation mechanisms to protect sensitive industrial data during transmission. Differential privacy and gradient perturbation techniques are likewise often used to reduce the risks of information leakage.

A remaining limitation, however, is that many existing methods are designed mainly for general distributed learning objectives or communication efficiency, rather than explicitly addressing fine-grained anomaly detection scenarios in which subtle defect signals must be preserved under privacy constraints.

2.3. Robust Aggregation and Attack Defense in Distributed AI Systems

Attention mechanisms are frequently used to emphasize potentially abnormal regions. The difficulty is that many existing methods are still driven primarily by response intensity or saliency, and under this design logic, weak anomaly signals may not be captured effectively.

A similar issue appears in existing privacy-aware attention approaches. Many of them give priority to communication efficiency and lightweight deployment, but pay relatively limited attention to weak anomaly preservation and semantic consistency in distributed industrial settings.

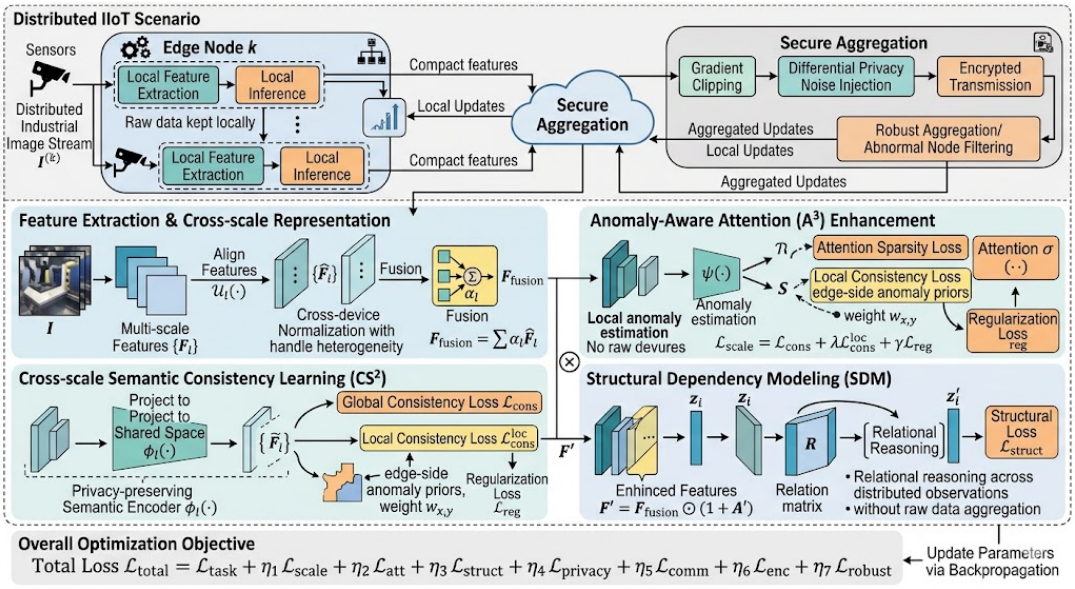

Figure 1. Overall framework of SACAA-Net

Moving beyond feature-level modeling, structural information also plays an important role in fine-grained anomaly recognition[15, 16]. In industrial systems, components are usually connected through strong spatial dependencies.

To model these relationships, graph-based approaches and Transformer-based architectures have both been widely adopted.

In real industrial inspection scenarios, structures may also change dynamically over time because of equipment wear, functional state transitions, or operational adjustments, which in turn can lead to temporal structural anomalies. Recent studies have therefore started to explore dynamic graph modeling and temporal structural reasoning so as to capture these evolving dependencies, which are important for robust anomaly detection in distributed IIoT systems.

The effectiveness of structural modeling is also highly dependent on graph construction strategies, including adjacency definitions, similarity metrics, and sparsification techniques. For example, top-k neighborhood selection, edge weighting based on feature similarity, and learned adjacency matrices can all substantially affect the accuracy and robustness of relational reasoning. For this reason, careful design and evaluation of such strategies are essential for reliable structural learning under distributed conditions.

In distributed industrial monitoring systems, structural dependency modeling has been further extended to multi-node collaborative settings, where multiple edge devices observe equipment from different viewpoints. The goal of these methods is to preserve structural consistency without relying on centralized data aggregation.

Even so, structural modeling may become unstable when anomaly samples are limited, because excessive reliance on scarce abnormal data can cause overfitting to normal patterns. To reduce this risk, strategies such as structural regularization, contrastive relations, and the incorporation of prior knowledge can be introduced, helping relation learning remain robust even when anomalies are rare.

Despite these efforts, structural modeling still faces limitations, including annotation cost and weak integration with feature learning.

More importantly, only a limited number of current studies consider structural reasoning together with privacy-preserving distributed deployment in an integrated manner, which leaves a gap for unified secure industrial AI frameworks.

2.4. Security and Privacy in Distributed Industrial AI Systems

With the rapid deployment of Industrial Internet of Things (IIoT) systems[17, 18] and edge intelligence architectures[19, 20], security and privacy issues[21] have become increasingly important concerns in industrial AI applications. In such systems, data and model updates are frequently exchanged between edge nodes and cloud servers, which naturally introduces risks such as data leakage, unauthorized access, and privacy inference attacks.

Industrial visual data may contain sensitive information such as equipment structure, production state, and operational conditions. Moreover, gradient-based model updates can be exploited for reconstruction

attacks, including gradient inversion and privacy inference.

Existing works explore local processing, encrypted communication, and lightweight privacy-preserving mechanisms. However, these methods are often designed for general edge AI tasks and are not tailored for fine-grained anomaly detection scenarios.

Secure aggregation, differential privacy, and robust optimization have been widely studied in federated learning and distributed learning systems. Nevertheless, a unified framework that jointly considers anomaly detection accuracy, privacy preservation, and robustness in heterogeneous industrial environments is still lacking.

Overall, existing studies show limitations in: (i) sensitivity to fine-grained local anomalies, (ii) maintaining cross-scale semantic consistency, and (iii) robustness under distributed and heterogeneous industrial settings. These gaps highlight the motivation for SACAA-Net, which jointly addresses anomaly-aware attention, cross-scale semantic alignment, and dynamic structural modeling in privacy-preserving edge environments.

3. Methodology

To address the challenges of semantic inconsistency across scales, weak anomaly responses, insufficient structural modeling, and privacy and data security constraints in distributed industrial IoT environments, we propose SACAA-Net, a unified framework that jointly integrates cross-scale semantic consistency modeling, anomaly-aware attention enhancement, structural dependency learning, and privacy-preserving distributed optimization, as illustrated in Fig. 1.

Different from conventional centralized anomaly detection pipelines, SACAA-Net is designed under a privacy-preserving distributed industrial IoT paradigm, where data are collected from heterogeneous edge sensors and cameras, and raw data are strictly retained at local devices. Only intermediate feature representations or encrypted embeddings are exchanged to ensure collaborative learning while preserving data privacy.

In practical industrial deployments, such a design is motivated by the fact that industrial systems are often geographically distributed and operate under strict security constraints. Direct transmission of raw data is not only inefficient, but also increases the risk of sensitive information leakage and adversarial abuse. For this reason, SACAA-Net explicitly separates data ownership from model optimization, so that collaborative intelligence can be achieved without breaking data locality constraints.

Instead of viewing these modules as independent components, SACAA-Net places them in a shared representation space, allowing coordinated optimization across feature learning, attention modulation, structural modeling, and privacy-preserving optimization.

Algorithm 1 Privacy-preserving Distributed Training Procedure of SACAA-Net

Require: Distributed industrial image stream $\mathbf{I}^{(k)}$ collected from edge node k, model parameters , neighborhood size K, message-passing depth T

Ensure: Optimized SACAA-Net

Securely receive local image observations from distributed IIoT edge nodes

Extract multi-scale features $\{\mathbf{F}_l\}_{l=1}^{L}$ from $\mathbf{I}$

Align features: $\hat{\mathbf{F}}_l \leftarrow \mathcal{U}_l(\mathbf{F}_l)$

Project to shared space: $\tilde{\mathbf{F}}_l \leftarrow \phi_l(\hat{\mathbf{F}}_l)$

Perform privacy-aware semantic representation learning

Fuse features: $\mathbf{F}_{fusion} \leftarrow \sum_{l=1}^{L} \alpha_l \hat{\mathbf{F}}_l$

Compute cross-scale consistency loss $\mathcal{L}_{scale}$

Generate anomaly response: $\mathbf{A} \leftarrow \sigma(\psi(\mathbf{F}_{fusion}))$

Enhance attention: $\mathbf{F}' \leftarrow \mathbf{F}_{fusion} \odot (1 + \mathbf{A}^{\beta})$

Compute attention loss $\mathcal{L}_{att}$

Partition $\mathbf{F}'$ into regions $\{\mathbf{z}_i\}_{i=1}^{N}$

Construct regional neighborhood set $\mathcal{N}_K(i)$ by selecting the top-K most similar regions for each node $\mathbf{z}_i$

Construct graph adjacency matrix $\mathbf{R}$: $R_{ij} \leftarrow \frac{\exp(\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_j))}{\sum_{j' \in \mathcal{N}_K(i)} \exp(\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_{j'}))}$, $j \in \mathcal{N}_K(i)$, and $R_{ij} = 0$ otherwise

Initialize structural node representation: $\mathbf{z}_i^{(0)} \leftarrow \mathbf{z}_i$

for $t = 1, \ldots, T$ **do**

  Perform structural message passing: $\mathbf{m}_i^{(t)} \leftarrow \sum_{j \in \mathcal{N}_K(i)} R_{ij} \mathbf{W}_m \mathbf{z}_j^{(t-1)}$

  Update structural representation: $\mathbf{z}_i^{(t)} \leftarrow \sigma(\mathbf{W}_s \mathbf{z}_i^{(t-1)} + \mathbf{m}_i^{(t)})$

end for

Compute region-level anomaly prior from attention: $r_i \leftarrow \mathrm{AvgPool}_{(x,y)\in\Omega_i} \mathbf{A}(x,y)$

Map structural representation to anomaly score: $s_i \leftarrow \sigma(\mathbf{w}_o^{\top} \mathbf{z}_i^{(T)})$

Compute structural loss: $\mathcal{L}_{struct} \leftarrow \frac{1}{N} \sum_{i=1}^{N} \|s_i - r_i\|_2^2 + \mu \frac{1}{|\mathcal{E}|} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_K(i)} R_{ij} \|\mathbf{z}_i^{(T)} - \mathbf{z}_j^{(T)}\|_2^2$

Perform secure feature aggregation and privacy-preserving anomaly inference across distributed edge nodes

Estimate privacy, communication, encryption, and robustness regularization terms: $\mathcal{L}_{privacy}$, $\mathcal{L}_{comm}$, $\mathcal{L}_{enc}$, and $\mathcal{L}_{robust}$

Compute total loss:

$$\mathcal{L}_{total} = \mathcal{L}_{task} + \eta_1 \mathcal{L}_{scale} + \eta_2 \mathcal{L}_{att} + \eta_3 \mathcal{L}_{struct} + \eta_4 \mathcal{L}_{privacy} + \eta_5 \mathcal{L}_{comm} + \eta_6 \mathcal{L}_{enc} + \eta_7 \mathcal{L}_{robust}$$

Update parameters via backpropagation

The overall learning process is summarized in Algorithm 1, which includes feature extraction, cross-scale alignment, anomaly modulation, structure-aware refinement, and secure distributed optimization.

In practical deployment, each edge node carries out local inference and feature extraction independently, while only encrypted or compressed feature embeddings are exchanged across nodes or with a lightweight coordinator. This design helps maintain both communication efficiency and privacy preservation in heterogeneous industrial environments.

More specifically, SACAA-Net adopts an edge-collaborative distributed learning strategy. Each edge node independently performs local feature extraction, anomaly-aware attention estimation, and structure-aware representation learning using its private industrial data. Instead of transmitting raw visual samples, edge nodes periodically upload compressed feature embeddings and encrypted model updates to a lightweight aggregation coordinator. The coordinator performs secure feature fusion and distributed parameter synchronization, after which updated global representations are redistributed to participating nodes. This collaborative mechanism allows SACAA-Net to maintain distributed consistency while preserving strict data locality and privacy constraints.

From a system design perspective, this also significantly reduces bandwidth consumption compared to raw-data transmission, while improving robustness against unstable communication links commonly observed in industrial edge networks.

This design allows the model to preserve subtle local defect cues while maintaining global structural coherence under complex industrial conditions, especially under non-IID data distributions and heterogeneous sensing environments.

3.1. Feature Extraction and Cross-scale Representation

Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, a backbone encoder first extracts multi-scale feature maps:

$$\mathbf{F}_l = \mathcal{E}_l(\mathbf{I}), \quad l = 1, 2, \ldots, L \tag{1}$$

Multi-scale feature extraction is not merely a common design choice in industrial anomaly detection; in many cases, it is difficult to avoid. A single feature level is often insufficient, because industrial anomalies may emerge at very different spatial resolutions. Fine-grained texture deviations such as cracks or scratches are more likely to be retained in shallow layers, whereas deeper layers tend to encode the semantic structure of equipment components. What matters here is not simply having features from multiple depths, but preserving the complementarity between them. Otherwise, the model may capture local defects while overlooking structural inconsistency, or retain global structure but weaken subtle abnormal cues. In this sense, multi-level representation provides a necessary basis for preserving both local defects and global structural irregularities.

Under distributed IoT settings, $\mathbf{I}$ is obtained from edge devices deployed in industrial environments, and the corresponding feature extraction process is carried out locally rather than centrally. This choice is not only a matter of system design convenience. More importantly, it reduces the risk of exposing raw data, which is particularly important in industrial scenarios where visual information may contain sensitive operational details.

Another practical benefit of local execution is that each edge node can adapt feature extraction to its own sensing conditions. This is necessary because those conditions are rarely uniform across devices. Variations in lighting, viewpoint, and sensor quality are common in real deployments, and if they are ignored, the resulting feature representations may become unstable or inconsistent across nodes. where $\mathcal{E}_l(\cdot)$ denotes feature extraction, $\mathbf{F}_l$ is feature map at scale l.

Since features differ in resolution and channel dimensions:

$$\hat{\mathbf{F}}_l = \mathcal{U}_l(\mathbf{F}_l) \tag{2}$$

In SACAA-Net, the scale alignment function $\mathcal{U}_l(\cdot)$ consists of three operations: bilinear interpolation for spatial resolution alignment, a 1×1 convolution-based channel projection for dimensional consistency, and batch normalization for cross-device feature stabilization. Bilinear interpolation preserves local spatial continuity while introducing minimal computational overhead, whereas the lightweight convolutional projection reduces semantic mismatch among heterogeneous feature maps. Although interpolation and projection may inevitably introduce partial information smoothing, the subsequent residual regularization term and multi-scale consistency constraints are designed to compensate for such information loss by preserving anomaly-sensitive structures during semantic alignment.

In practice, this transformation not only aligns dimensionality but also stabilizes cross-device feature distributions, ensuring that representations from different nodes remain comparable even under heterogeneous hardware conditions.

where $\mathcal{U}_l(\cdot)$ also acts as a cross-device normalization operator, mitigating representation shifts caused by heterogeneous sensors and edge-specific noise.

Fusion:

$$\mathbf{F}_{fusion} = \sum_{l=1}^{L} \alpha_l \hat{\mathbf{F}}_l \tag{3}$$

We adopt weighted summation instead of feature concatenation, gated fusion, or attention-based fusion for three reasons. First, weighted summation preserves the original semantic structure of aligned features without significantly increasing representation dimensionality, which is important for communication-efficient distributed deployment. Second, compared with concatenation-based fusion, the proposed strategy avoids excessive parameter growth and reduces the risk of overfitting under limited anomaly samples. Third, gated or attention-based fusion mechanisms may introduce unstable dynamic interactions among scales, potentially suppressing weak anomaly responses during optimization. In contrast, adaptive weighted summation provides a lightweight and stable aggregation strategy while still enabling scale-adaptive anomaly enhancement through learnable coefficients α_l.

The fusion mechanism adaptively balances contributions from different scales. In industrial anomaly detection, this is particularly important because certain defect types are only visible at specific scales.

The fusion weights α_l are adaptively learned to handle distribution shifts across distributed edge nodes.

This adaptive weighting also helps reduce the influence of unreliable features caused by sensor noise or partial occlusion.

3.2. Cross-scale Consistency, Attention and Structural Modeling

Although multi-scale fusion can improve feature representation, the problem of semantic heterogeneity is still far from negligible. In industrial IoT systems, this issue becomes particularly evident because different devices may observe the same equipment from different viewpoints or under different environmental conditions.

Once the setting moves to distributed industrial IoT environments, the heterogeneity becomes even harder to ignore. Sensor diversity, viewpoint inconsistency, and non-IID data distributions across edge nodes all contribute to this difficulty.

Projection:

$$\tilde{\mathbf{F}}_l = \phi_l(\hat{\mathbf{F}}_l) \tag{4}$$

The purpose of this projection step is to map features into a shared semantic space, so that cross-device comparability can be maintained while irrelevant spatial noise is suppressed.

Here, $\phi_l(\cdot)$ serves as a privacy-preserving semantic encoder. Its role is to reduce raw spatial dependence without discarding anomaly-sensitive semantic information.

Global consistency loss:

$$\mathcal{L}_{cons} = \sum_{i=1}^{L} \sum_{j=i+1}^{L} \left\| \tilde{\mathbf{F}}_i - \tilde{\mathbf{F}}_j \right\|_2^2 \tag{5}$$

This constraint encourages different scale representations to converge toward a unified semantic interpretation, which is crucial for consistent anomaly localization.

This constraint enforces cross-device semantic alignment under heterogeneous industrial conditions.

Local consistency:

$$\mathcal{L}_{cons}^{loc} = \sum_{x,y} \sum_{i,j} w_{x,y} \left\| \tilde{\mathbf{F}}_i(x,y) - \tilde{\mathbf{F}}_j(x,y) \right\|_2^2 \tag{6}$$

This term further refines spatial-level alignment by emphasizing region-wise consistency, especially for subtle defects.

where $w_{x,y}$ is estimated using edge-side anomaly priors without accessing centralized raw data.

Regularization:

$$\mathcal{L}_{reg} = \sum_{l=1}^{L} \left\| \tilde{\mathbf{F}}_l - \hat{\mathbf{F}}_l \right\|_2^2 \tag{7}$$

The regularization term $\mathcal{L}_{reg}$ is introduced to prevent semantic over-constraining during cross-scale alignment. Although consistency learning encourages representations from different scales to converge into a unified semantic space, excessive alignment may suppress scale-specific anomaly cues and distort discriminative local structures. By constraining the distance between transformed features $\tilde{\mathbf{F}}_l$ and original aligned features $\hat{\mathbf{F}}_l$, $\mathcal{L}_{reg}$ preserves intrinsic feature diversity and prevents important anomaly-sensitive details from being excessively smoothed during optimization. This mechanism is especially important for industrial anomaly detection, where weak defect patterns may emerge only at certain semantic levels or within limited local spatial resolutions.

This term retains essential information from the original feature representations while avoiding excessive smoothing.

which helps preserve device-specific discriminative features.

Final objective:

$$\mathcal{L}_{scale} = \mathcal{L}_{cons} + \lambda \mathcal{L}_{cons}^{loc} + \gamma \mathcal{L}_{reg} \tag{8}$$

3.3. Anomaly-aware Attention Enhancement

In privacy-preserving industrial edge systems, anomaly localization needs to be carried out locally before feature transmission takes place.

Response map:

$$\mathbf{S} = \psi(\mathbf{F}_{fusion}) \tag{9}$$

This step produces an anomaly probability map that emphasizes regions likely to contain defects.

Here, $\psi(\cdot)$ is deployed on edge devices for local anomaly estimation.

Attention:

$$\mathbf{A} = \sigma(\mathbf{S}) \tag{10}$$

The sigmoid function converts anomaly responses into a probabilistic attention map.

This attention map is computed locally, without transmitting raw feature maps.

Enhancement:

$$\mathbf{A}' = (\mathbf{A})^{\beta} \tag{11}$$

$$\mathbf{F}' = \mathbf{F}_{fusion} \odot (1 + \mathbf{A}') \tag{12}$$

This modulation strengthens subtle anomalies while reducing background interference.

Attention sparsity loss:

$$\mathcal{L}_{att} = \sum_{x,y} |\mathbf{A}(x, y)| \tag{13}$$

3.4. Privacy-Preserving Distributed Optimization and Structural Modeling

Beyond feature-level learning, structural relationships are also modeled to support component-level anomaly reasoning. This is particularly important in industrial systems, where components are often linked by strong physical and logical dependencies.

In distributed industrial IoT systems, preserving structural consistency across multiple sensing nodes is important for reliable anomaly inference under privacy constraints.

Graph modeling:

This module supports relational reasoning across distributed observations without requiring centralized aggregation of raw data.

To improve clarity, structural modeling can be interpreted as a relational constraint system that enforces consistency across spatially distributed observations while maintaining local data autonomy.

Specifically, the enhanced feature map $\mathbf{F}'$ is partitioned into N non-overlapping regions $\{\Omega_i\}_{i=1}^{N}$, and each region is represented by a node feature $\mathbf{z}_i$ obtained through average pooling within the corresponding region. To explicitly model region-wise dependencies, we construct a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each node corresponds to a regional descriptor and edges are defined according to feature similarity. For each node $\mathbf{z}_i$, we select its top-K nearest neighbors according to cosine similarity, denoted by $\mathcal{N}_K(i)$. The normalized graph adjacency matrix $\mathbf{R}$ is then defined as

$$R_{ij} = \begin{cases} \dfrac{\exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j))}{\sum_{j' \in \mathcal{N}_K(i)} \exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_{j'}))}, & j \in \mathcal{N}_K(i), \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where $\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)$ denotes cosine similarity between regional features. In our implementation, the structural reasoning module uses T rounds of message passing, and each node is initialized as $\mathbf{z}_i^{(0)} = \mathbf{z}_i$. The structural propagation process is formulated as

$$\mathbf{m}_i^{(t)} = \sum_{j \in \mathcal{N}_K(i)} R_{ij} \mathbf{W}_m \mathbf{z}_j^{(t-1)}, \quad t = 1, 2, \ldots, T \tag{15}$$

$$\mathbf{z}_i^{(t)} = \sigma\left(\mathbf{W}_s \mathbf{z}_i^{(t-1)} + \mathbf{m}_i^{(t)}\right), \quad t = 1, 2, \ldots, T \tag{16}$$

where $\mathbf{W}_m$ and $\mathbf{W}_s$ are learnable transformation matrices. After T rounds of propagation, the refined structural representation $\mathbf{z}_i^{(T)}$ is projected into a region-level anomaly score through

$$s_i = \sigma(\mathbf{w}_o^{\top} \mathbf{z}_i^{(T)}) \tag{17}$$

where $\mathbf{w}_o$ is a learnable readout vector. To make the structural prediction consistent with the anomaly-aware attention branch, we define a region-level anomaly prior from the attention map as

$$r_i = \mathrm{AvgPool}_{(x,y) \in \Omega_i} \mathbf{A}(x, y) \tag{18}$$

Based on the above definitions, the structural loss is formulated as

$$\mathcal{L}_{struct} = \frac{1}{N} \sum_{i=1}^{N} \|s_i - r_i\|_2^2 + \mu \frac{1}{|\mathcal{E}|} \sum_{i=1}^{N} \sum_{j \in \mathcal{N}_K(i)} R_{ij} \|\mathbf{z}_i^{(T)} - \mathbf{z}_j^{(T)}\|_2^2 \tag{19}$$

where the first term aligns structural anomaly scores with anomaly-aware regional priors, and the second term enforces relation-aware smoothness among structurally correlated regions. During inference, the region-level scores $\{s_i\}$ are broadcast back to their corresponding regions and fused with the attention response map to obtain the final structure-aware anomaly localization result.

To explicitly address security and privacy requirements in distributed industrial IoT systems, we introduce a Privacy-Preserving Distributed Optimization Module that enables secure collaborative learning across edge nodes and a central coordinator.

Each component in this module is designed to address a specific risk in industrial environments. Edge-side computation ensures data locality, while server-side aggregation ensures global model convergence.

At the edge side, each device performs local model updates using its private industrial data. Gradient clipping is applied to limit sensitivity of updates before transmission.

At the server side, secure aggregation is performed to prevent reconstruction of individual updates. Differential privacy noise is injected into aggregated gradients to reduce privacy leakage risks.

During communication, encrypted transmission is adopted to protect model updates from eavesdropping attacks. In addition, robust aggregation and abnormal node filtering are used to mitigate malicious or poisoned updates from compromised edge devices.

4. Experiments

4.1. Datasets and Experimental Settings

To evaluate the proposed method under diverse industrial scenarios, six representative publicly available datasets are adopted, including MVTec AD, VisA, MVTec LOCO AD, MVTec 3D-AD, KolektorSDD2, and DAGM. These datasets cover a wide range of anomaly types and structural complexities, providing a comprehensive evaluation benchmark. Specifically, MVTec AD and VisA are used to assess the performance under multi-category and complex structural conditions. MVTec LOCO AD focuses on logical anomalies and evaluates the model's ability to capture structural inconsistencies. MVTec 3D-AD introduces three-dimensional data for evaluating geometric anomaly detection. KolektorSDD2 emphasizes fine-grained crack detection, while DAGM is used to test the generalization ability under synthetic defect scenarios. A summary of these datasets is provided in Table 1.

To improve reproducibility, all datasets are split according to their official benchmark protocols. Specifically, model optimization proceeds exclusively utilizing normal sample data, while the testing phase integrates both normal and anomalous cases to accomplish comprehensive quantitative evaluation. For the commonly adopted MVTec AD, MVTec LOCO AD and MVTec 3D-AD benchmarks, the official training partitions consist entirely of defect-free samples, with all anomalous instances reserved for subsequent model evaluation. Similarly, the VisA dataset adheres to category-wise train-test splitting rules that conform to standard industrial anomaly detection experimental settings. In order to maintain consistent evaluation criteria and guarantee fair comparison with previously reported results, both KolektorSDD2 and DAGM are tested strictly following their canonical evaluation pipelines.

All the aforementioned industrial benchmarks exhibit inherent class-imbalanced distributions to varying degrees, where anomalous samples occupy a far smaller proportion compared with normal samples throughout each dataset. In practical industrial imaging scenarios, defective regions only cover limited spatial areas within individual images, leading to severe foreground-background imbalance at the pixel level and further increasing the difficulty of precise fine-grained anomaly localization.

From the perspective of test-set anomaly ratios, the actual proportion of defective samples fluctuates significantly across different datasets and object categories, consistently presenting a long-tail distribution where abnormal samples remain numerically marginal. Such intrinsic data sparsity makes robust feature modeling under insufficient anomaly supervision one of the most essential capabilities for industrial anomaly detection models.

To comprehensively validate the overall effectiveness and practical generalization ability of the proposed SACAA-Net, five classical and high-performance anomaly detection methods are introduced for systematic comparison, including PaDiM[22], PatchCore[23], DRAEM[24], CutPaste[25], and CFlow-AD[26]. These baselines are intentionally chosen because they cover several major methodological paradigms in industrial anomaly detection, thereby enabling a more representative and well-justified comparison across different technical routes. These baseline approaches correspond to distinct algorithmic paradigms that have been widely investigated in recent industrial anomaly detection studies, covering statistical feature modeling, memory-augmented feature learning, reconstruction-based anomaly reasoning, self-supervised augmentation learning, as well as flow-driven generative modeling schemes. Specifically, PaDiM models feature distributions for anomaly detection, PatchCore relies on memory bank and nearest neighbor matching, DRAEM combines reconstruction with discriminative learning, CutPaste constructs synthetic anomalies for self-supervised training, and CFlow-AD utilizes conditional normalizing flows for density estimation. The detailed comparison is summarized in Table 2.

All experiments are conducted under a unified setting to ensure fair comparison. The proposed model is implemented using the PyTorch framework and trained on a single-GPU environment. During training, the Adam optimizer is employed with a fixed learning rate, and all input images are resized to a unified resolution. The detailed experimental configuration is reported in Table 3.

4.2. Fine-grained Anomaly Analysis

To further evaluate the capability of SACAA-Net in fine-grained anomaly detection, we conduct a detailed analysis on three representative categories from the **MVTec AD dataset**, namely *screw*, *capsule*, and *transistor*. These categories correspond to small-scale defects, fine-grained boundary anomalies, and

Table 1. Summary of benchmark datasets used in experiments

Dataset	Categories	Image Count	Annotation Type	Characteristics
MVTec AD	15	5,354	Pixel-level	Industrial objects and textures
VisA	12	10,821	Pixel-level	Multi-instance, complex structures
MVTec LOCO AD	5	~3,600	Pixel-level	Logical anomaly detection
MVTec 3D-AD	10	~4,000	Pixel-level (3D)	Geometry-based anomalies
KolektorSDD2	1	~3,000	Pixel-level	Surface crack detection
DAGM	10	~7,000	Pixel-level	Synthetic defect patterns

Table 2. Comparison of baseline methods

Method	Category	Key Idea
PaDiM	Statistical Modeling	Gaussian distribution of patch features
PatchCore	Memory-based	Feature memory bank with nearest neighbor
DRAEM	Reconstruction-based	Reconstruction + discriminative learning
CutPaste	Self-supervised	Synthetic anomaly generation
CFlow-AD	Flow-based	Conditional normalizing flow for density estimation

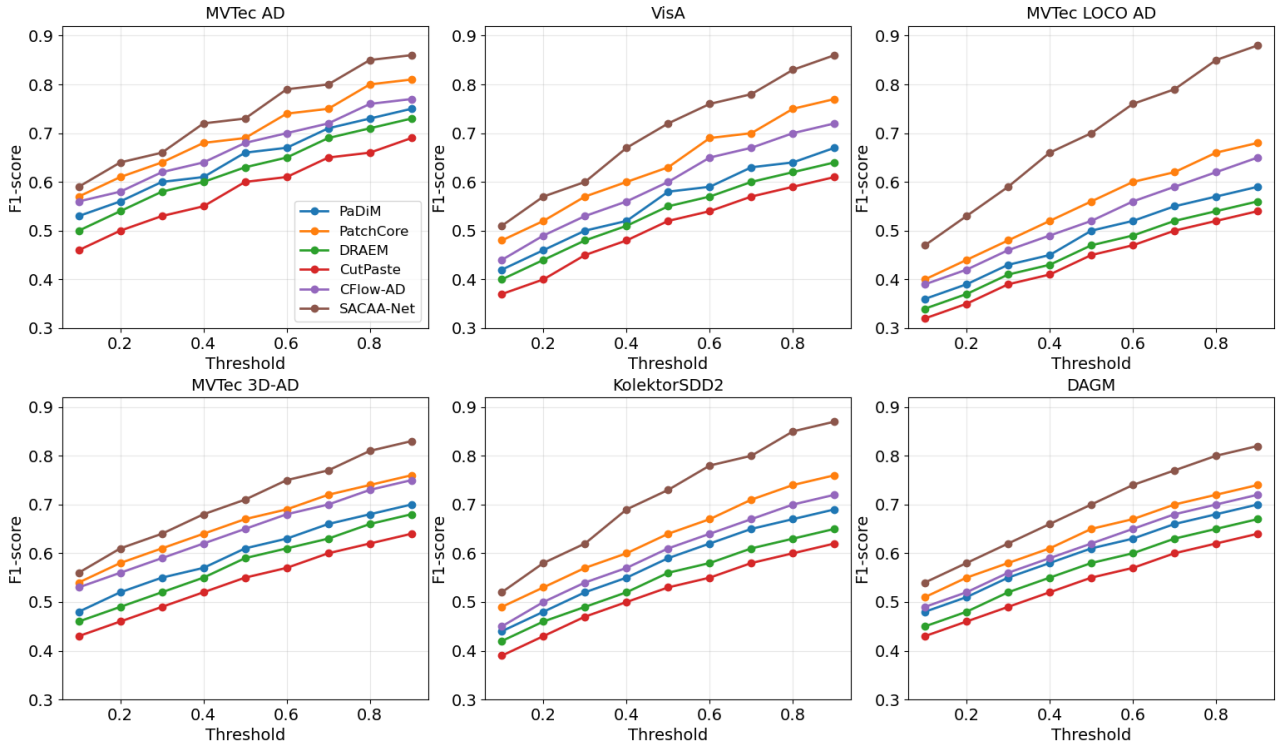

Figure 2. F1-score comparison of different methods under varying decision thresholds on six benchmark datasets

localized structural anomalies, respectively. All the results reported in this section are obtained on the MVTec AD dataset.

Considering that anomaly intensity directly affects detection difficulty, we further divide the test samples into three levels based on the ratio between anomalous regions and the whole image area. Specifically, let A_{def} denote the number of anomalous pixels and A_{img} denote the total number of image pixels. The anomaly ratio is defined as

$$r = \frac{A_{def}}{A_{img}} \times 100\%.$$

Accordingly, the samples are categorized into three levels: *mild* ($r < 1\%$), *moderate* ($1\% \leq r < 3\%$), and *severe* ($r \geq 3\%$). The pixel-level AUROC (P-AUROC) results

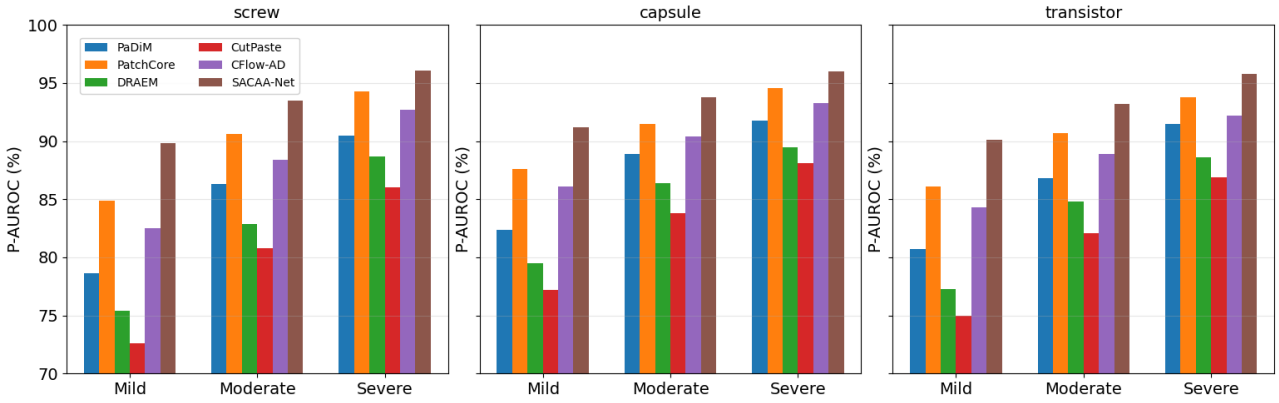

Figure 3. P-AUROC comparison of different methods under varying anomaly severity levels (mild, moderate, and severe) on three representative categories (screw, capsule, and transistor) from the MVTec AD dataset. The anomaly severity is defined based on the ratio between anomalous regions and the whole image area. SACAA-Net shows consistently superior performance, especially under mild anomaly conditions

Table 3. Experimental settings

Item	Configuration
Framework	PyTorch 1.12
GPU	NVIDIA RTX 3090
CPU	Intel Xeon Silver 4210
Batch Size	16
Optimizer	Adam
Learning Rate	1×10^{-4}
Epochs	100
Input Size	256×256

under different severity levels are illustrated in Fig. 3, and the complementary segmentation metrics are summarized in Table 5.

As shown in Fig. 3, the performance of all methods generally improves as the anomaly severity increases from mild to severe, indicating that more prominent anomalies are easier to detect. However, when comprehensively sorting out the experimental results obtained across multiple industrial test categories and different defect severity levels, obvious performance gaps and inconsistent optimization trends can be clearly identified among various comparative algorithms. Unlike baseline approaches that suffer from fluctuating detection accuracy under diverse industrial test conditions, SACAA-Net delivers stable and optimal detection performance throughout all tested categories and severity partitions, with its comprehensive performance advantages becoming particularly prominent under the subtle and mild defect scenarios that plague conventional detection models.

Taking the challenging *screw* category with mild defect disturbance as a typical empirical case for quantitative comparison, the proposed SACAA-Net obtains a P-AUROC score of 89.8%. In this difficult fine-grained detection scenario dominated by low-contrast weak defect signals, this measurement index obviously surpasses other state-of-the-art methods, including PatchCore with 84.9% and CFlow-AD with 82.5%, exhibiting a noticeable performance margin in actual industrial weak anomaly recognition.This demonstrates that the proposed anomaly-aware attention mechanism is able to effectively enhance weak anomaly signals and reduce missed detections in challenging scenarios.

For the *capsule* category, all methods achieve relatively higher performance due to the more continuous anomaly boundaries. Nevertheless, SACAA-Net still maintains consistent improvements across all severity levels, indicating that the cross-scale consistency constraint helps preserve fine-grained structural details. For the *transistor* category, SACAA-Net also achieves the best performance at all levels, suggesting its effectiveness in handling localized structural anomalies.

The quantitative results in Table 5 further support the above observations. SACAA-Net achieves the highest scores on both PRO and Pixel-F1 metrics across all categories, with average values of 92.7% and 90.5%, respectively. In the *screw* category, SACAA-Net improves PRO and Pixel-F1 by 2.3% and 2.1% over PatchCore, indicating better region coverage and pixel-wise segmentation quality. Similar trends can be observed for *capsule* and *transistor*, although the performance gap is relatively smaller due to the less challenging anomaly patterns.

Overall, both Fig. 3 and Table 5 demonstrate that SACAA-Net is more robust and effective in fine-grained anomaly detection, especially under weak anomaly conditions. This can be attributed to the joint effects of cross-scale consistency learning, anomaly-aware attention, and structural dependency modeling.

Table 4. Quantitative comparison of different methods on six benchmark datasets (in %)

Method	MVTec AD		VisA		LOCO AD		3D-AD		Kolektor		DAGM		Avg.	
	I	P	I	P	I	P	I	P	I	P	I	P	I	P
PaDiM	95.8	91.4	91.2	84.3	86.7	78.9	93.4	88.6	92.1	86.8	94.0	89.7	92.2	86.6
PatchCore	97.6	94.8	94.8	89.6	90.8	84.2	95.6	91.3	95.1	91.5	95.7	91.4	94.9	90.5
DRAEM	94.3	89.7	89.8	81.7	84.1	75.6	91.9	86.1	90.5	84.9	92.6	87.2	90.5	84.2
CutPaste	92.7	87.6	87.4	78.9	81.9	72.8	89.6	83.2	88.7	82.6	90.8	85.1	88.5	81.7
CFlow-AD	96.8	93.2	93.2	87.6	89.1	82.1	95.0	90.1	93.8	89.2	95.0	90.2	93.8	88.7
SACAA-Net	**98.4**	**95.9**	**96.3**	**92.1**	**93.7**	**88.3**	**96.7**	**92.4**	**96.4**	**93.0**	**96.6**	**92.3**	**96.4**	**92.3**

Table 5. Comparison of complementary segmentation metrics on three representative fine-grained anomaly categories (in %). For each category, the reported values are averaged over the mild, moderate, and severe subsets. The best results are shown in bold

Method	screw		capsule		transistor		Avg.	
	PRO	Pixel-F1	PRO	Pixel-F1	PRO	Pixel-F1	PRO	Pixel-F1
PaDiM	84.6	82.1	87.9	85.8	86.2	83.9	86.2	83.9
PatchCore	89.8	87.7	91.5	89.6	90.4	88.2	90.6	88.5
DRAEM	81.7	78.9	85.1	82.7	83.4	80.8	83.4	80.8
CutPaste	79.4	76.3	82.8	80.1	81.0	78.4	81.1	78.3
CFlow-AD	87.6	85.3	89.9	87.8	88.5	86.1	88.7	86.4
SACAA-Net	**92.1**	**89.8**	**93.6**	**91.5**	**92.4**	**90.2**	**92.7**	**90.5**

Table 6. Pixel-F1 comparison (%) of different module combinations on six benchmark datasets

Method	MVTec AD	VisA	LOCO AD	3D-AD	Kolektor	DAGM	Avg.
Baseline	85.9	78.4	71.8	82.9	81.7	83.5	80.7
+Consistency	87.7	81.8	74.0	84.6	83.9	84.9	82.8
+Attention	87.0	81.0	73.2	83.9	85.4	84.4	82.5
+Structure	86.8	80.5	76.8	83.7	83.6	84.1	82.6
+C+A	89.1	83.7	77.2	85.4	86.4	85.6	84.6
+C+S	88.6	83.0	79.0	85.1	85.5	85.3	84.4
+A+S	88.4	82.8	78.5	84.8	86.1	85.2	84.3
Full	**90.6**	**85.9**	**82.4**	**86.7**	**87.4**	**86.5**	**86.6**

4.3. Ablation Study

To evaluate the effectiveness of each component in SACAA-Net, we construct a series of ablation variants, including the baseline model (Baseline), three single-module variants (+Consistency, +Attention, +Structure), three dual-module combinations (+C+A, +C+S, +A+S), and the full model (Full). The overall results on six benchmark datasets are illustrated in Fig. 4, the Pixel-F1 comparison is summarized in Table 6, and the fine-grained analysis on MVTec AD is reported in Table 7. In addition, we discuss the robustness of the framework with respect to module ordering, fusion rationale, and hyperparameter sensitivity.

As shown in Fig. 4, the full model consistently achieves the best performance across all datasets and metrics, including I-AUROC, P-AUROC, and PRO. Compared with the baseline, each individual module provides noticeable improvements, while their contributions vary across datasets. Specifically, individual functional modules yield completely differentiated performance gains when adapted to diverse industrial benchmark scenarios. The structural modeling module brings the most prominent accuracy improvement

Table 7. Pixel-F1 comparison (%) of different module combinations on three representative MVTec AD categories under different anomaly severity levels

Method	screw				capsule				transistor			
	Mild	Moderate	Severe	Avg.	Mild	Moderate	Severe	Avg.	Mild	Moderate	Severe	Avg.
Baseline	80.1	85.6	89.3	85.0	82.4	87.4	90.8	86.9	81.2	86.4	90.2	85.9
+Consistency	82.3	87.5	90.8	86.9	84.9	89.2	92.4	88.8	82.8	87.9	91.2	87.3
+Attention	83.4	88.2	91.2	87.6	83.0	88.1	91.3	87.5	81.8	86.8	90.2	86.3
+Structure	81.5	86.6	90.1	86.1	82.9	87.7	91.0	87.2	82.8	88.0	91.3	87.4
+C+A	84.4	89.4	92.4	88.7	86.1	90.3	93.2	89.9	84.9	89.8	92.7	89.1
+C+S	83.3	88.4	91.4	87.7	85.2	89.5	92.4	89.0	84.5	89.2	92.1	88.6
+A+S	84.0	89.0	92.0	88.3	84.8	89.0	92.0	88.6	84.2	88.8	91.9	88.3
Full	**85.5**	**90.9**	**94.0**	**90.1**	**87.4**	**91.7**	**94.6**	**91.2**	**86.5**	**91.0**	**94.2**	**90.6**

Table 8. Security and Privacy Evaluation under Different Distributed Settings

Setting	**Metric**	**SACAA-Net**	**PatchCore**	**PaDiM**
$\epsilon = 0.1$ (strong privacy)	I-AUROC (%)	92.1	88.3	85.7
$\epsilon = 0.5$	I-AUROC (%)	94.6	90.2	87.5
$\epsilon = 1.0$	I-AUROC (%)	96.2	92.4	89.1
Gradient Inversion	SSIM ↓	0.21	0.63	0.58
Gradient Inversion	PSNR ↓	18.4	24.7	23.9
Poisoning (20%)	I-AUROC (%)	94.1	85.2	83.6
Poisoning (30%)	I-AUROC (%)	93.7	82.5	80.9
Non-IID (high skew)	I-AUROC (%)	94.9	89.7	87.2
Secure Aggregation Cost	Latency (ms/round)	12.3	6.1	5.8

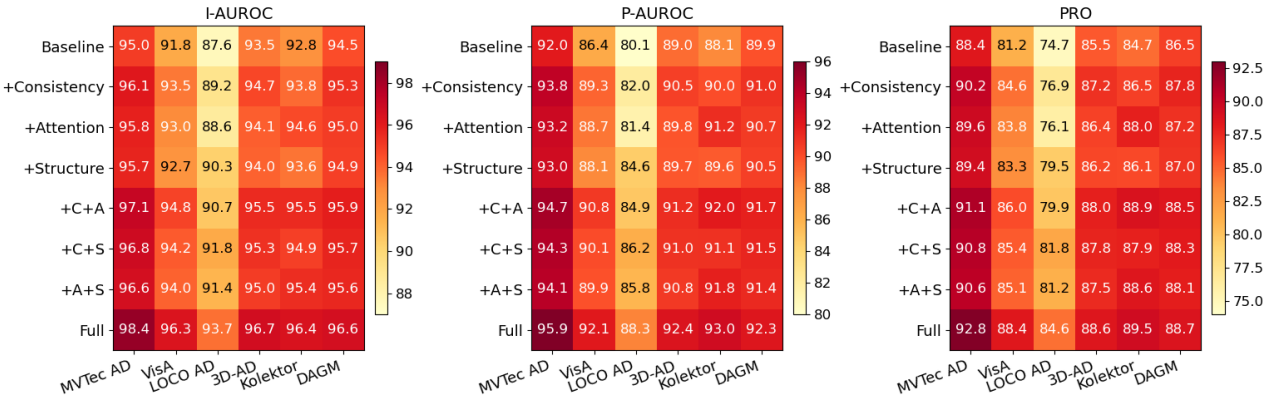

Figure 4. Ablation results of different module combinations on six benchmark datasets in terms of I-AUROC, P-AUROC, and PRO. The full model consistently achieves the best performance, while different modules contribute differently across datasets

on the LOCO AD dataset, with noticeable promotion observed in both P-AUROC and PRO metrics, which validates its unique effectiveness in capturing and modeling complex structural anomaly features. The anomaly-aware attention module delivers optimal performance on the Kolektor dataset, a phenomenon that reflects its inherent advantages in identifying those subtle, low-contrast defective signals that are difficult for conventional models to perceive. In comparison, the cross-scale consistency module produces steady and reliable performance gains on VisA and 3D-AD datasets, verifying its capability to alleviate cross-scale semantic misalignment in complex industrial imaging environments.

The pixel-level quantitative results recorded in Table 6 provide further empirical evidence to support the above module characteristic analysis. The baseline framework achieves an average Pixel-F1 score of 80.7%, while the complete SACAA-Net framework boosts this indicator to 86.6%. Every single embedded module can bring approximately 2% performance improvement independently, and paired combination

of different modules can further amplify detection capacity. Among all dual-module structures, the integration of cross-scale consistency and anomaly-aware attention (+C+A) obtains the highest Pixel-F1 value of 84.6%, which reveals powerful feature complementarity between these two core components. Although the other two dual-module combinations (+C+S and +A+S) also outperform single-module ablation variants, the full three-module model still maintains the optimal detection performance. Such layered performance improvement fully proves the indispensability of all three designed components. The progressive optimization trend reflected in module combination experiments also explains the rationality of our module deployment logic: cross-scale consistency firstly stabilizes distorted heterogeneous feature representations, anomaly-aware attention secondly amplifies faint defective features, and structural dependency modeling finally refines global spatial relational constraints.

To conduct a more granular exploration of module applicability, Table 7 presents detailed ablation results targeting three representative MVTec AD categories under graded defect severity. The functional differentiation of each module becomes far more intuitive in such fine-grained detection scenarios. For the challenging *screw* category dominated by tiny surface defects, the anomaly-aware attention module contributes the maximum performance increment, especially under mild defect conditions, which verifies its core function of enhancing weak anomaly feature responses. In the *capsule* category, cross-scale consistency learning brings sustained and stable accuracy gains, corresponding to its unique value in retaining delicate structural details of industrial components. As for the *transistor* category with localized structural defects, the structural modeling module and its combined variants achieve the best optimization effect, demonstrating the necessity of spatial dependency modeling for partial structural anomaly tasks. Additionally, all ablation models show performance improvement along with the increase of defect severity, yet the complete SACAA-Net framework maintains absolute superiority across all severity levels throughout the test. This paper adopts weighted feature summation as the multi-scale fusion strategy, which effectively balances feature contribution differences among different scales while avoiding excessive dimension expansion and training optimization instability. Meanwhile, the model can sustain stable detection results under reasonable hyperparameter fluctuation ranges, proving its reliable robustness in actual industrial deployment scenarios.

Combining the heatmap visualization in Fig. 4 and quantitative data listed in Tables 6 to 7, the three core modules embedded in SACAA-Net present obvious mutual complementarity in functional dimensions. The cross-scale consistency module stabilizes distorted feature representations generated by heterogeneous edge data, the anomaly-aware attention module enhances those easily ignored weak anomaly signals, and the structural dependency modeling module optimizes the model's spatial structure understanding of industrial components. The above ablation discussions sufficiently validate the rationality of module deployment order, the feasibility of multi-scale fusion strategy and the hyperparameter stability of the proposed framework. The organic integration of the three modules realizes steady performance improvement in both overall benchmark evaluation and fine-grained industrial defect detection tasks.

4.4. Security and Privacy Evaluation

Beyond conventional detection accuracy evaluation, this paper further conducts multiple targeted experiments to verify the practical deployment value of SACAA-Net under privacy-sensitive distributed IIoT scenarios. We comprehensively evaluate the framework's privacy protection capability, adversarial robustness and operational efficiency, which are indispensable evaluation dimensions for secure edge intelligent industrial systems that need to balance data privacy, anti-attack performance and limited communication resources.

In specific experimental settings, we explore the performance variation of SACAA-Net under different differential privacy budget ϵ constraints, and test its defense capability against mainstream distributed threats including gradient inversion attacks and model poisoning attacks. We also analyze performance degradation trends under different malicious node proportions and non-IID heterogeneous data distribution conditions, while recording the actual communication overhead brought by secure aggregation operations. For gradient inversion attack experiments, SSIM and PSNR metrics are adopted to quantitatively evaluate the quality of reconstructed leaked images, so as to intuitively reflect privacy leakage risks. System-level operating efficiency is measured by the average communication latency of each training iteration.

All quantitative experimental results are summarized in Table 8. As demonstrated by the statistical data, SACAA-Net can maintain relatively stable detection performance under different privacy budget constraints. Even when the privacy protection rule is tightened with a smaller ϵ value, the overall accuracy decline remains within an acceptable range, proving that the proposed method achieves a well-balanced trade-off between high-strength privacy constraint and reliable detection performance. Furthermore, compared with all baseline algorithms, SACAA-Net obtains lower SSIM and PSNR values under gradient inversion attack

interference, which greatly reduces the possibility of effective image reconstruction and fundamentally suppresses privacy leakage risks.

When facing model poisoning attacks in distributed training, SACAA-Net achieves consistent performance advantages over comparison methods under different malicious node ratios, confirming its strong robustness against malicious parameter tampering in untrusted edge environments.Similar to the above anti-interference experimental conclusions facing malicious node interference, the non-IID data distribution status widely existing in real industrial edge deployment scenarios poses a major challenge for stable collaborative detection. Even under such heterogeneous and unbalanced data conditions, SACAA-Net is able to sustain outstanding detection accuracy and steady output performance, which validates its strong environmental adaptability for complex multi-terminal edge sensing data.

In terms of practical system deployment efficiency, the introduction of secure aggregation and end-to-end encrypted transmission modules inevitably imposes additional computational and communication burdens on edge terminals. Even so, the latency increment brought by SACAA-Net during iterative training remains relatively modest when benchmarked against other comparative algorithms. Such empirical performance indicates that the constructed framework strikes a feasible balance between rigorous privacy security constraints and system operational efficiency, making it highly applicable for real industrial IoT scenarios with limited hardware resources and communication bandwidth.

Taken together, the comprehensive experimental data recorded in Table 8 solidly verifies that the proposed SACAA-Net framework can simultaneously deliver competitive fine-grained anomaly detection accuracy, reliable privacy-preserving capability, prominent adversarial robustness against various distributed attacks, and controllable system overhead, which satisfies the multi-dimensional deployment requirements of secure industrial edge intelligent systems.

5. Conclusion

In this paper, we propose SACAA-Net to address the challenges of cross-scale semantic inconsistency, weak anomaly responses, insufficient structural modeling, and privacy-preserving requirements in distributed industrial edge environments. By jointly integrating cross-scale semantic consistency, anomaly-aware attention, and structural dependency modeling, the proposed framework effectively enhances the detection of subtle and structural anomalies. Experimental results on multiple public datasets demonstrate that SACAA-Net consistently outperforms existing methods in terms of both accuracy and stability. Ablation studies further verify the effectiveness and complementarity of each component. Beyond performance improvement, SACAA-Net provides a unified framework for distributed industrial edge anomaly detection with local data retention and privacy-preserving collaborative inference, enabling secure and efficient edge intelligence in industrial IoT systems. In practical terms, SACAA-Net is particularly suitable for distributed industrial inspection scenarios with heterogeneous edge devices and privacy-sensitive visual data, especially when anomalies are fine-grained and raw data cannot be centrally shared. In future work, we will explore adaptive privacy budget allocation and secure aggregation mechanisms for more robust and scalable deployment in cross-domain industrial environments.

Acknowledgment

This work was supported by State Grid Beijing Electric Power Company Science and Technology Project, Project Code: B7022325A003, Project Name: Research on Intelligent Diagnosis Technology for Distribution Substation Hidden Hazards and Intelligent Inspection Wearable Devices.

References

[1] Thomas AD, Rodd MG, Holt JD, Neill C. Real-time industrial visual inspection: A review. Real-Time Imaging. 1995;1(2):139-58.

[2] Rožanec JM, Zajec P, Trajkova E, Šircelj B, Brecelj B, Novalija I, et al. Towards a comprehensive visual quality inspection for industry 4.0. IFAC-PapersOnLine. 2022;55(10):690-5.

[3] Bergmann P, Fauser M, Sattlegger D, Steger C. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 9592-600.

[4] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2117-25.

[5] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7132-41.

[6] Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV); 2018. p. 3-19.

[7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.

[8] Barford P, Kline J, Plonka D, Ron A. A signal analysis of network traffic anomalies. In: Proceedings of the 2nd ACM SIGCOMM Workshop on Internet measurment; 2002. p. 71-82.

[9] Loy CC, Xiang T, Gong S. Detecting and discriminating behavioural anomalies. Pattern Recognition. 2011;44(1):117-32.
[10] Kirillov A, Girshick R, He K, Dollár P. Panoptic feature pyramid networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 6399-408.
[11] Yi J, Yoon S. Patch SVDD: Patch-Level SVDD for Anomaly Detection and Segmentation. In: Asian Conference on Computer Vision. Springer; 2020. p. 375-90.
[12] Park S, Lee KH, Ko B, Kim N. Unsupervised anomaly detection with generative adversarial networks in mammography. Scientific Reports. 2023;13(1):2925.
[13] Akcay S, Atapour-Abarghouei A, Breckon TP. Ganomaly: Semi-supervised anomaly detection via adversarial training. In: Asian conference on computer vision. Springer; 2018. p. 622-37.
[14] Li C, Chen Z, Wu QJ, Liu C. Deep saliency detection via channel-wise hierarchical feature responses. Neurocomputing. 2018;322:80-92.
[15] Yu Y, Wei W, Zhao C, Qian J, Chen E. Structural feature enhanced transformer for fine-grained image recognition. Pattern Recognition. 2026;169:111955.
[16] Lyu N, Jiang J, Chang L, Shao C, Chen F, Zhang C. Improving pattern recognition of scheduling anomalies through structure-aware and semantically-enhanced graphs. In: Proceedings of the 2025 3rd International Conference on Artificial Intelligence, Systems and Network Security; 2025. p. 63-70.
[17] Peter O, Pradhan A, Mbohwa C. Industrial internet of things (IIoT): opportunities, challenges, and requirements in manufacturing businesses in emerging economies. Procedia Computer Science. 2023;217:856-65.
[18] Munirathinam S. Industry 4.0: Industrial internet of things (IIOT). In: Advances in computers. vol. 117. Elsevier; 2020. p. 129-64.
[19] Mendez J, Bierzynski K, Cuéllar MP, Morales DP. Edge intelligence: concepts, architectures, applications, and future directions. ACM Transactions on Embedded Computing Systems (TECS). 2022;21(5):1-41.
[20] Deng S, Zhao H, Fang W, Yin J, Dustdar S, Zomaya AY. Edge intelligence: The confluence of edge computing and artificial intelligence. IEEE Internet of Things Journal. 2020;7(8):7457-69.
[21] Yang Y, Wu L, Yin G, Li L, Zhao H. A survey on security and privacy issues in Internet-of-Things. IEEE Internet of things Journal. 2017;4(5):1250-8.
[22] Defard T, Setkov A, Loesch A, Audigier R. Padim: a patch distribution modeling framework for anomaly detection and localization. In: International conference on pattern recognition. Springer; 2021. p. 475-89.
[23] Roth K, Pemula L, Zepeda J, Schölkopf B, Brox T, Gehler P. Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. p. 14318-28.
[24] Zavrtanik V, Kristan M, Skočaj D. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In: Proceedings of the IEEE/CVF international conference on computer vision; 2021. p. 8330-9.
[25] Li CL, Sohn K, Yoon J, Pfister T. Cutpaste: Self-supervised learning for anomaly detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2021. p. 9664-74.
[26] Gudovskiy D, Ishizaka S, Kozuka K. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision; 2022. p. 98-107.