IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 January 2026, accepted 8 February 2026, date of publication 3 March 2026, date of current version 10 March 2026.

*Digital Object Identifier 10.1109/ACCESS.2026.3670056*

RESEARCH ARTICLE

# Self-Attention Enhanced Student–Teacher Framework for Visual Anomaly Detection in Industrial Images

**HAFIZ A. AMJAD[1], IRIDA SHALLARI[2], (Member, IEEE), MATTIAS O' NILS[2], (Member, IEEE), AND FAISAL Z. QURESHI[1], (Senior Member, IEEE)**

[1]Faculty of Science, University of Ontario Institute of Technology, Oshawa, ON L1G OC5, Canada
[2]Department of Computer and Electrical Engineering, Mid Sweden University, 85170 Sundsvall, Sweden

Corresponding author: Faisal Z. Qureshi (faisal.qureshi@ontariotechu.ca)

This work was supported in part by the National Sciences and Engineering Research Council of Canada (NSERC) Discovery under Grant RGPIN 005159/2020, and in part by the Mitacs Accelerate under Grant IT17268.

**ABSTRACT** In this work we introduce AeCSAD, a student–teacher framework designed to detect two types of anomalies in images of industrial components: structural anomalies (physical defects) and logical anomalies (incorrect relationships between components). Unlike prior methods, AeCSAD extends the component segmentation–based logical anomaly detection scheme (c. 2024) with self-attention mechanisms, enabling more effective relational modeling. We demonstrate consistent improvements on the MVTec LOCO Anomaly Detection benchmark. Specifically, AeCSAD employs a global student network with self-attention for reasoning across distant components, complemented by a local student network for fine-grained analysis. Additionally, a patch histogram module measures the frequency distribution of components, allowing the system to detect irregularities in their occurrence. During inference, anomaly scores from the histogram module and the fused local–global networks are combined to produce the final anomaly score. Experiments show that AeCSAD achieves superior average AUROC performance on both structural and logical anomaly detection tasks compared to prior approaches.

**INDEX TERMS** Industrial visual inspection, logical anomaly detection, self-attention mechanism.

## I. INTRODUCTION

Manufacturing industries underpin modern economies [1], yet they remain vulnerable to unpredictable factors such as material variation, assembly errors, foreign object intrusion, and wear [2], [3]. These issues manifest as visual anomalies, which are visible deviations from expected appearance and degrade manufacturing performance, safety, and reliability. Broadly, such visual anomalies can fall into two categories: structural anomalies, which involve disruptions in geometric properties such as shape, texture, or alignment; and logical anomalies, which arise when semantic relationships between objects or their components violate contextual expectations, even though the individual parts appear correct [4], [5].

Prior to the adoption of deep learning, visual anomaly detection relied on classical statistical models that characterized normal data distributions in a compact feature space. These methods include Principal Component Analysis (PCA) [6], One-Class Support Vector Machines (SVM) [7], and Gaussian Mixture Models (GMMs) [8], which detect anomalies as deviations in distance, density, or reconstruction error from the normal data manifold. Statistical techniques were particularly effective for detecting anomalies where low-level deviations in texture, edge structure, or brightness could be measured reliably. However, these methods exhibited high sensitivity to variations in lighting, viewpoint, and noise, and they typically required extensive manual tuning of features and parameters. Such limitations motivated a shift toward data-driven architectures capable of learning representations directly from raw inputs.

The associate editor coordinating the review of this manuscript and approving it for publication was Andrea F. Abate.

Data-driven methods for visual anomaly detection initially focused on localized appearance variations, leveraging Convolutional Neural Networks (CNNs) [9] to capture spatial features through fixed receptive fields [10], [11]. A receptive field refers to the specific region of an image that a CNN processes at a given layer. This localized processing imparts CNNs with an inductive bias toward capturing local spatial patterns, making them particularly effective for detecting structural anomalies [10], [11]. This design has led to strong performance on structural anomaly benchmarks such as MVTec AD [12], [13], [14], [15], often exceeding 99% AUROC [13], [16]. However, CNNs remain limited in modeling logical anomalies [4], [17], which require reasoning across spatially distant and semantically related components. As a result, CNN-based architectures may fail to detect arrangements that appear locally correct but violate global compositional consistency. To address this, several methods have introduced segmentation-aware pipelines that approximate logical reasoning [18], [19]. For example, Component-Aware Anomaly Detection (ComAD) [18] segments images into pseudo-parts using unsupervised clustering and flags anomalies based on deviations in part statistics such as size or location. Similarly, Part Segmentation-based Anomaly Detection (PSAD) [19] uses few-shot part annotations to construct logic-aware memory banks encoding expected component layouts. Although these methods introduce logic-awareness, they do so indirectly by relying on statistical heuristics or feature similarity, without explicitly modeling relational semantics in the architecture.

A more structured formulation of logical anomaly detection emerges in Component Segmentation Anomaly Detection (CSAD) framework [20] which performs semantic part segmentation using Recognize Anything Model++ (RAM++) [21] to generate open-vocabulary tags, GroundingDINO [22] to localize component bounding boxes, and Segment Anything Model (SAM) [23] to extract component masks. The component masks are clustered to define pseudo-classes for training a segmentation network, which enables histogram-based analysis to detect inconsistencies in component count and layout. Moreover, CSAD has a Local-Global Student-Teacher (LGST) module consisting of a frozen teacher network, a local student, and a global student. The local student is trained to replicate the teacher's features using a limited receptive field of $33 \times 33$, making it sensitive to localized structural anomalies. In contrast, the global student processes the entire image through a CNN-based autoencoder that compresses spatial information into a $1 \times 1$ embedding [24]. Although, the global student in CSAD covers the entire image through a CNN-based autoencoder with a $1 \times 1$ bottleneck, it remains limited in its ability to capture long-range dependencies and relational structures. This is because standard CNNs, even with large receptive fields, rely on hierarchical local feature aggregation and lack explicit mechanisms to model interactions between spatially distant components [25], [26]. As a result, the global student in CSAD approximates logical consistency using appearance-based surrogates rather than modeling relational semantics explicitly.

To address this limitation, we propose Attention Enhanced CSAD (AeCSAD), which augments the global student network in CSAD with self-attention blocks [26]. Self-attention explicitly captures distant relational dependencies across images, improving the ability of the global student for modeling component-level relationships and global context. Moreover, this architectural enhancement integrates naturally with CSAD's student-teacher framework. Since the teacher defines the normal distribution of features, the student must not only replicate appearance but also infer the spatial and semantic relationships encoded in the teacher's outputs. Self-attention amplifies the student's sensitivity to such structure, enabling it to fail more meaningfully in the presence of logically inconsistent arrangements. This self-attention based enhancement forms the principal architectural contribution of this work and lead us to the following objectives for developing AeCSAD:

1) **To develop a unified, lightweight anomaly detection framework** capable of addressing both structural and logical anomalies under unsupervised conditions;
2) **To design a modular architecture** that separates local appearance from relational reasoning of components;
3) **To integrate both pixel-level sensitivity and semantic-level understanding** within a single framework that balances local detail capture with global context modeling.

To achieve these objectives, we propose Attention Enhanced CSAD (AeCSAD), an anomaly detection framework that augments CSAD with self-attention in the global student network. Our main contributions are as follows:

- We propose AeCSAD, an anomaly detection framework that augments the CSAD architecture by integrating self-attention blocks into the global student network to enable explicit relational reasoning across distant image regions.
- AeCSAD jointly addresses both structural and logical anomalies by combining both local and global student networks with a histogram-based module that captures component frequency and layout statistics.
- Extensive experiments on the MVTec LOCO AD benchmark demonstrate that AeCSAD outperforms state-of-the-art methods in terms of AUROC across both structural and logical anomaly categories.

## II. RELATED WORK

A large body of work in visual anomaly detection is built on convolutional features combined with non-parametric or statistical mechanisms to model normal features. Methods such as SPADE [27], PaDiM [15], and RegAD [28] score anomalies by comparing test-time features to those drawn from normal data, using either k-nearest neighbors, Gaussian modeling, or few-shot registration. These approaches capture structural deviations effectively, leveraging local appearance cues and spatial alignment to detect

surface-level defects. Building on this foundation, PatchCore [13] and its successors, including PNI [29], ReConPatch [30], and InReaCh [31], refine kNN-based detection pipelines to improve memory efficiency, robustness, and spatial sensitivity. These enhancements introduce elements such as coreset sampling, positional priors, contrastive feature refinement, and outlier filtering, expanding the operational range of retrieval-based detectors. SimpleNet [32] further streamlines appearance-based detection by adapting pretrained CNN features with a lightweight feature adaptor and training a discriminator on Gaussian-perturbed normal features, offering a fast, memory-efficient alternative to retrieval-based methods. While these methods are effective at identifying localized anomalies, they primarily operate within an appearance-similarity paradigm and lack architectural mechanisms for explicit semantic reasoning.

Another prominent line of work leverages the Student–Teacher (ST) paradigm, where a student model is trained to replicate the output of a frozen teacher on normal data, with deviations serving as anomaly signals. Uninformed Students [33] use ensembles of student regressors over multiscale CNN features, with error variance indicating anomaly regions. Reverse Distillation [34] introduces architectural asymmetry and a one-class embedding bottleneck to increase anomaly sensitivity, while Asymmetric Student–Teacher (AST) [35] further strengthens this paradigm by explicitly decoupling teacher and student capacities through a bijective normalizing-flow teacher and a lightweight CNN student, preventing undesired generalization and improving anomaly separability. MemKD [36] augments this with memory-guided normality recall to stabilize the student's representations. GAP [37], though not a classical ST model, compares local patches with their surrounding context via dual-head discriminators, capturing spatial inconsistencies through relational mismatches.

A complementary direction centers on modeling object-level or part-level structure. Component-aware Anomaly Detection (ComAD) [18] segments images into pseudo-components using unsupervised ViT features and evaluates their spatial and statistical consistency. PSAD [19] introduces few-shot supervised segmentation to model part distributions and inter-component relationships via multi-source memory fusion. EfficientAD [38], in contrast, achieves logic-aware detection without segmentation by pairing student–teacher outputs with an auxiliary autoencoder, whose reconstruction failure reveals global inconsistencies. These methods significantly improve reasoning about component presence and arrangement, but are constrained either by segmentation quality (ComAD), label dependence (PSAD), or indirect approximations of logical structure (EfficientAD).

Collectively, these methods represent substantial progress toward logic-aware anomaly detection. Yet many retain disjoint pipelines that separate structural and logical cues, rely on auxiliary memory modules or segmentation heads, or defer reasoning to inference-time heuristics. Moreover, their logic-centric modeling remains largely implicit and focus on reconstruction or feature alignment rather than symbolic reasoning or compositional structure. In contrast, AeCSAD unifies low-level structural modeling and high-level compositional reasoning within a single student–teacher framework, enhanced with self-attention. By embedding relational understanding directly into the global student's training objective, AeCSAD offers an end-to-end architecture that captures spatial, semantic, and structural dependencies.

## III. METHODOLOGY

### A. OVERVIEW OF THE PROPOSED FRAMEWORK

Our proposed method, Attention-Enhanced CSAD (AeCSAD), extends the CSAD framework [20], which is organized around three core modules. The first module discussed in sections III-B, III-C, and III-D forms the semantic pseudo-label generation and training pipeline. This module generates component-level segmentation masks using RAM++, GroundingDINO, and SAM. These masks are then clustered to distinguish a high-confidence labeled subset and low-confidence unlabeled subset which are then used to train a lightweight component-level segmentation network consisting of a WideResNet-50 based encoder and DeepLABv3+ based decoder. The second module, called Patch Histgoram, discussed in Section III-E, captures component frequencies for detecting count and layout irregularities. The third module called the Local-Global Student-Teacher, discussed in Section III-F, includes a local student and global student network. These students are trained to mimic a frozen teacher model and support detection of structural and logical inconsistencies. We preserve this modular design of CSAD and introduce a self-attention mechanism in global student to capture long-range inter-component dependencies, enhancing the detection of logical anomalies. Figure 1 presents a high-level conceptual overview of AeCSAD.

### B. SEMANTIC PSEUDO-LABELED MASK GENERATION

The semantic pseudo-labeled mask generation process begins by passing a normal image through RAM++ [21], which generates textual tags for both objects and background. These tags are filtered following the procedure described in CSAD [20] where non-object or irrelevant background tags are removed to ensure semantic relevance. The resulting tags are then used to prompt GroundingDINO [22], which outputs corresponding bounding boxes for object and background regions. Figure 2 illustrates the conceptual pipeline of RAM++ for generating textual tags, while Figure 3 depicts the GroundingDINO process for generating object and background bounding boxes. These bounding boxes then guide the first run of SAM, producing coarse object-level masks $\mathcal{M}_{\mathrm{G}}$. To extract finer component-level details, a second run of SAM is executed in automatic mode (i.e., without prompts), yielding a dense set of candidate masks $\mathcal{M}_{\mathrm{S}}$. Figure 4 shows a conceptual illustration

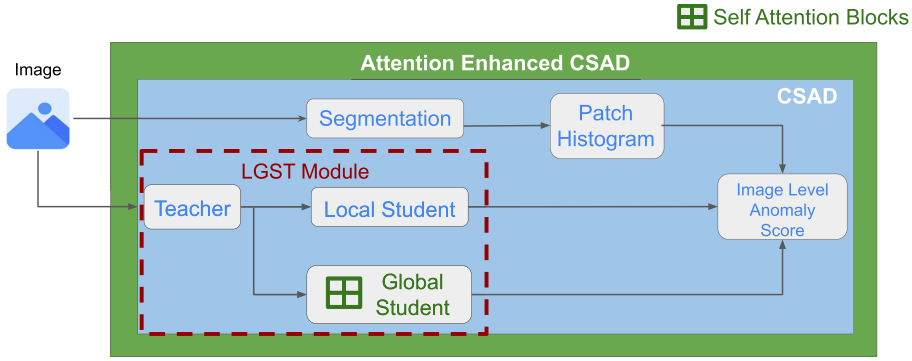



**FIGURE 1.** Conceptual illustration of AeCSAD architecture.

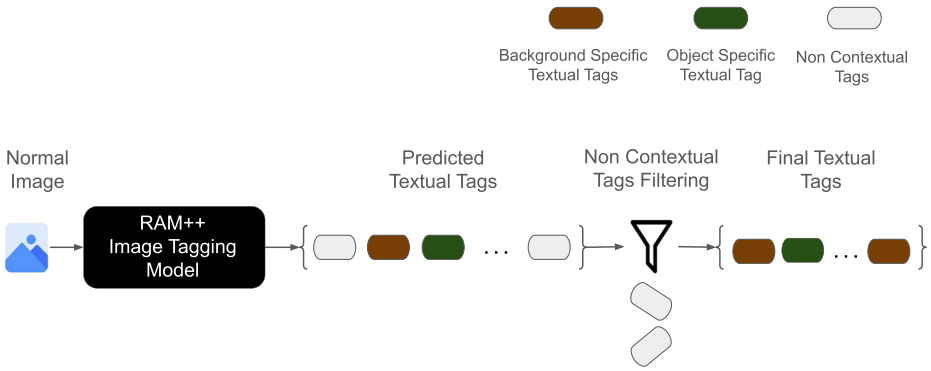



**FIGURE 2.** Conceptual illustration of textual tags generation using RAM++.

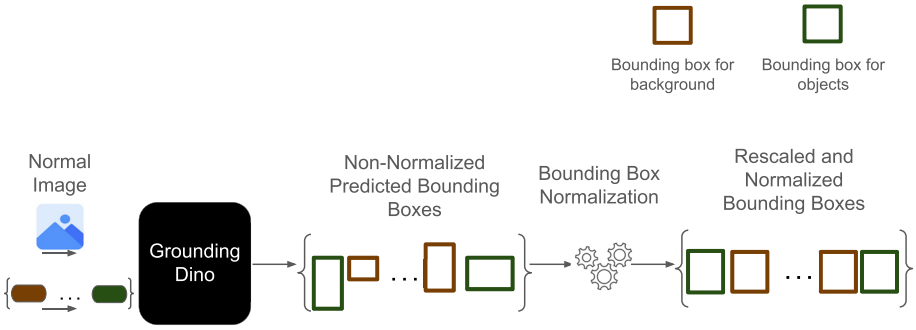



**FIGURE 3.** Conceptual illustration of background and object bounding box generation using GroundingDINO.

of component-level mask generation using SAM. A two-stage intersection-based filtering is then applied to refine these masks that performs (1) background intersection filtering to remove any mask in $\mathcal{M}_S$ that significantly overlaps with background regions in $\mathcal{M}_G$, and (2) object intersection filtering to retain only those masks that have high intersection with object regions in $\mathcal{M}_G$.

To ensure that components with similar shape and texture are assigned the same pseudo label, it is important to minimize the impact of rotational variation. To achieve this, each retained mask $m_i \in \mathcal{M}^*$ is used to extract a square crop from the original image, enclosing the component region. This step results in a set of image patches $\mathcal{P} = \{p_i\}$, where each $p_i$ corresponds to a component that undergoes dense rotational augmentation across 60 angles uniformly spaced from 0° to 360°. Each rotated patch is then resized to $64 \times 64$ and passed through a pretrained WideResnet-50 encoder $\phi(\cdot)$, which extracts features from the 4th layer. These features are aggregated to form a fixed-length representation $f_i = \phi(p_i) \in \mathbb{R}^d$. The resulting set of vectors for each image forms

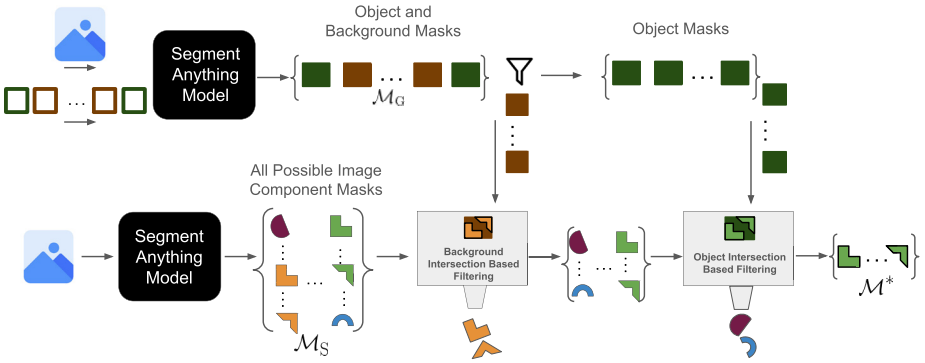



**FIGURE 4.** Component-level mask generation using SAM. Stage one uses bounding boxes and prompts to produce object and background masks, forming the set $\mathcal{M}_G$ after removing background. Stage two runs SAM in automatic mode to generate candidate masks $\mathcal{M}_S$, which are refined through (i) background intersection filtering and (ii) object alignment with $\mathcal{M}_G$, yielding the final mask set $\mathcal{M}^*$.

a feature matrix $F^{(i)} = [f_1, f_2, \ldots, f_{n^{(i)}}]^T$, where $n^{(i)}$ denotes the number of components in the $i$th image.

### C. PSEUDO-LABELED COMPONENT CLUSTERING

To assign pseudo labels to segmented components, the MeanShift algorithm [39] is applied to the rotation-invariant feature embeddings to cluster visually similar components without requiring the number of clusters to be specified in advance. After clustering, each patch $p_i$ is assigned a cluster label $c_i \in \{1, 2, \ldots, C\}$, where $C$ is the number of classes. Here, $C$ is not manually chosen. Instead, it emerges automatically from the density modes identified by MeanShift in feature space. MeanShift groups patches by locating local maxima of the kernel density estimate, and each density mode corresponds to one visually coherent component type. As a result, the number of pseudo-classes $C$ is fully data-driven and adapts to the structural complexity of each MVTec LOCO AD category. Categories with a small variety of part appearances typically yield fewer clusters, whereas categories with richer part variability produce a larger $C$. This allows the framework to infer semantically meaningful component types without predefined labels or prior knowledge of part counts. Each cluster is then mapped back onto the corresponding segmented region, yielding a pseudo-class index for each component. These assignments are then projected back onto their original spatial locations to form dense, full-resolution semantic masks $Y \in \mathbb{N}^{H \times W}$, which supervise the segmentation model during training. To evaluate pseudo-label reliability, class histograms are computed per image, representing each image's compositional structure in a compact form. HDBSCAN [40] is applied to this histogram space to identify clusters of images with similar class distributions. Images falling outside dominant histogram clusters due to over- or under-segmentation artifacts are considered unreliable. This two-stage clustering partitions the training set into (i) a high-confidence labeled subset $\mathcal{D}_L = \{(x_i^{(l)}, y_i^{(l)})\}$, and (ii) a low-confidence unlabeled subset $\mathcal{D}_U = \{x_j^{(u)}\}$.

### D. COMPONENT-LEVEL SEGMENTATION NETWORK TRAINING

A component-level segmentation network, consisting of a WideResNet-50-based frozen encoder pretrained on ImageNet [41], [42] and a DeepLabV3+ decoder with Atrous Spatial Pyramid Pooling (ASPP) [43] is trained exclusively on $\mathcal{D}_L$ and $\mathcal{D}_U$. The encoder extracts multi-scale features from intermediate layers, which are fed into a decoder where ASPP captures contextual dependencies at multiple receptive fields, producing dense features that balance global context and spatial precision.

The model is initially trained on using supervised losses on $\mathcal{D}_L$ using Cross-Entropy loss $\mathcal{L}_{\text{CE}}$, Dice loss $\mathcal{L}_{\text{Dice}}$, and Focal loss $\mathcal{L}_{\text{Focal}}$. Once stable representations are learned, auxiliary signals from $\mathcal{D}_U$ are introduced through entropy minimization $\mathcal{L}_{\text{Entropy}}$ and histogram alignment loss $\mathcal{L}_{\text{hist}}$. Following PSAD [19], $\mathcal{L}_{\text{hist}}$ is used to enforce part-level consistency by aligning the class-wise pixel volumes of unlabeled predictions with those of randomly selected labeled masks. This staged training strategy helps mitigate overfitting to noisy pseudo-labels and enhances generalization. The segmentation training loss is a weighted sum of all five objectives and is computed as

$$\mathcal{L}_{\text{total}}^{\text{seg}} = \lambda_1 \mathcal{L}_{\text{CE}} + \lambda_2 \mathcal{L}_{\text{Dice}} + \lambda_3 \mathcal{L}_{\text{Focal}} + \lambda_4 \mathcal{L}_{\text{hist}} + \lambda_5 \mathcal{L}_{\text{Entropy}}, \tag{1}$$

where $\lambda_1, \ldots, \lambda_5$ are empirically set as $[1.0, 0.5, 0.5, 0.3, 0.1]$. It is important to clarify that this segmentation loss $\mathcal{L}_{\text{total}}^{\text{seg}}$ is used *exclusively* to train the component-level segmentation network. To support real-time inference and embedded deployment, RAM++, GroundingDINO and SAM are not used during test time. Instead, each image is directly processed through the trained DeepLabV3+ decoder to produce a component-wise semantic segmentation map. These maps are subsequently passed to the Patch Histogram module.

### E. PATCH HISTOGRAM MODULE

The Patch Histogram module uses two modes (i) component-frequency modeling that primarily focuses on frequency

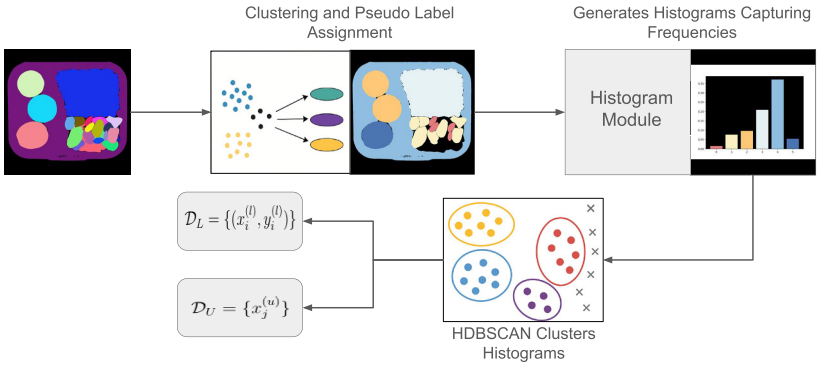



**FIGURE 5.** Overview of the clustering and histogram pipeline. Pseudo-label masks are clustered into semantic classes, followed by histogram extraction to capture component distributions.

of component classes and (ii) component-layout modeling that primarily focuses on layout of component classes where each mode is tailored to different types of anomalies. For component-frequency modeling, a reference Gaussian distribution is estimated using high confidence pseudo label subset $\mathcal{D}_L$ to calculate the empirical mean $\mu_{\text{Hist}}$ and covariance $\Sigma_{\text{Hist}}$. Given a predicted segmentation map $\hat{Y} \in \mathbb{N}^{H\times W}$, where each pixel is assigned a cluster index, its $\text{Hist}(\hat{Y}) \in \mathbb{R}^C$ is computed and its deviation from the referenced distribution is calculated using equation

$$\mathcal{L}_{\text{class}} = \sqrt{(\text{Hist}(\hat{Y}) - \mu_{\text{Hist}})^T \Sigma_{\text{Hist}}^{-1} (\text{Hist}(\hat{Y}) - \mu_{\text{Hist}})}. \quad (2)$$

For anomalies where the component frequency is the same as of normal image but the position of component is disrupted, component-layout modeling is employed. Here a segmentation map $\hat{Y}$,is partitioned into a $P \times P$ spatial grid. Each patch in this grid defines a localized region $B_{ij}$, where $i, j \in \{1, \ldots, P\}$. For each bin $B_{ij}$, a histogram vector $h_{ij} \in \mathbb{R}^C$ is calculated, where $C$ denotes the number of pseudo-classes. Each entry $h_{ij}^c$ encodes the number of pixels in region $B_{ij}$ that are assigned to component class $c$. These local histograms $h_{ij} \in \mathbb{R}^C$ are concatenated in raster-scan order to form a global layout vector $H \in \mathbb{R}^{P^2 \cdot C}$, encoding the layout of components across image. During testing, the normal image is passed through the same process to produce $\text{Hist}(Y)$, and its alignment with the test prediction $\text{Hist}(\hat{Y})$ is evaluated using hist loss $\mathcal{L}_{\text{hist}}$, employed for consistency regularization in segmentation training [19], [20]. Here, $\mathcal{L}_{\text{hist}}$ is repurposed to measure distributional deviation between predicted and reference component layouts using

$$\mathcal{L}_{\text{hist}} = \frac{1}{C} \sum_{c=1}^{C} \left\| \text{Hist}(\hat{Y}[c]) - \text{Hist}(Y[c]) \right\|. \quad (3)$$

Both $\mathcal{L}_{\text{class}}$ and $\mathcal{L}_{\text{hist}}$ function as distributional deviation measures evaluated on predicted segmentation maps. These quantities provide image-level scores reflecting how the component frequencies and spatial arrangements of a test sample differ from the learned normal reference. Their role is limited to inference, as the only histogram-related supervision used during training is the alignment regularizer $\mathcal{L}_{\text{hist}}$ included in the segmentation loss of Section III-D.

### F. SELF-ATTENTION ENHANCED AUTOENCODER IN AeCSAD

AeCSAD retains the structure of Local-Global Student-Teacher (LGST) module from CSAD but introduces self-attention blocks in the global student. Figure 6 shows a conceptual illustration of anomaly map generation in AeCSAD. The teacher network is a WideResNet-50 pretrained on ImageNet, which extracts multi-scale feature maps from intermediate layers. These features are spatially aligned and projected to produce the reference feature map $F_T$. The local student $\mathcal{S}_{\text{local}}$ is a shallow CNN with a $33 \times 33$ receptive field, designed to be sensitive to localized structural anomalies. It outputs a per-pixel feature map $F_L$, which is trained to match features of teacher using the mean squared error (MSE) loss using

$$\mathcal{L}_{\text{local}} = \frac{1}{BCHW} \sum_{i,j} \|F_T(i,j) - F_L(i,j)\|_2^2. \quad (4)$$

The global student $\mathcal{S}_{\text{global}}$ is constructed as an encoder-decoder autoencoder interleaved with self-attention layers. Its encoder consists of six convolutional layers that progressively reduce the spatial resolution while increasing channel dimensionality, culminating in a compact latent representation. Self-attention blocks are inserted after the fourth and fifth encoder layers, operating at feature resolutions of $16 \times 16$ and $8 \times 8$, respectively. These blocks apply scaled dot-product attention using

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V, \quad (5)$$

where $Q, K, V$ are linear projections of input tokens, and $d_k$ is the scaling factor. Each attention block reshapes spatial feature maps into sequences, applies attention, and restores the spatial structure through residual fusion. The decoder reconstructs the latent features using eight convolutional layers interleaved with bilinear upsampling operations. Two additional self-attention blocks are applied

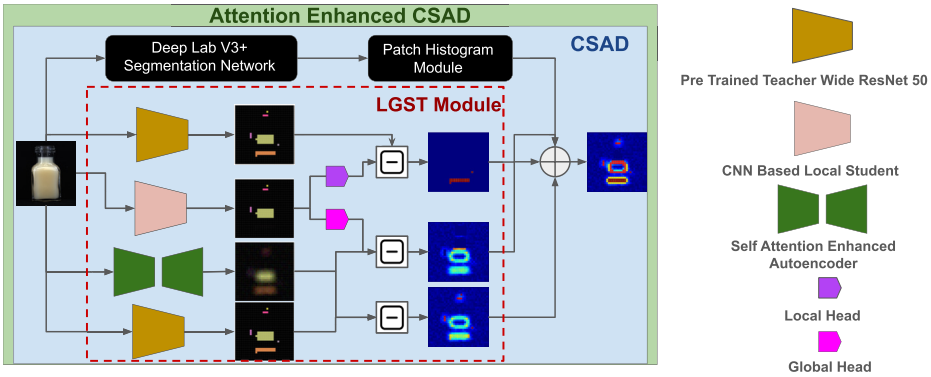



**FIGURE 6.** Conceptual illustration of anomaly map generation in AeCSAD. The LGST module comprises a frozen teacher network $\mathcal{T}_{\text{teacher}}$, based on a pretrained WideResNet-50 encoder; a CNN-based local student $\mathcal{S}_{\text{local}}$; and a global student $\mathcal{S}_{\text{global}}$, implemented as a self-attention enhanced autoencoder. Both student branches attempt to replicate the teacher's feature representations, and their deviations are used to construct multi-scale anomaly maps. The segmentation network and Patch Histogram module support semantic-guided anomaly detection downstream.

after the third and fifth decoder layers, at resolutions of $15 \times 15$ and $63 \times 63$, respectively, to promote long-range dependency modeling during upsampling. The global student produces a reconstructed feature map $F_G^{\text{attention}}$, which is supervised via:

$$\mathcal{L}_{\text{global}} = \frac{1}{BCHW} \sum_{i,j} \left\| F_T(i,j) - F_G^{\text{attention}}(i,j) \right\|_2^2. \quad (6)$$

Beyond the architectural modifications introduced by the self-attention layers, the student–teacher formulation itself plays a central role in enabling the detection of logical anomalies. The frozen teacher, pretrained on large-scale image corpora, produces feature activations that encode object structure, part–to–part relationships, and global compositional regularities. When the local and global student networks are trained to mimic these teacher features using only normal samples, they are implicitly guided to internalize the teacher's notion of structural and semantic coherence. Logical violations disrupt the global patterns encoded by the teacher, causing the global student with its self-attention layers to fail to reproduce the teacher's responses even when local appearance remains intact. This yields a characteristic feature mismatch that is most pronounced in categories with rich relational structure (e.g., *pushpins*, *screw bag*). This mechanism underlies the effectiveness of the LGST design for logical anomaly detection and motivates the integration of self-attention in AeCSAD. To quantify the discrepancy between the local-only features and the global attention-enhanced representation, the auxiliary fusion loss is used as:

$$\mathcal{L}_{\text{fusion}} = \frac{1}{BCHW} \sum_{i,j} \left\| F_G^{\text{attention}}(i,j) - F_L(i,j) \right\|_2^2. \quad (7)$$

The total objective for the LGST module is therefore defined as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{local}} \mathcal{L}_{\text{local}} + \lambda_{\text{global}} \mathcal{L}_{\text{global}} + \lambda_{\text{fusion}} \mathcal{L}_{\text{fusion}}, \quad (8)$$

where $\lambda_{\text{local}}, \lambda_{\text{global}}, \lambda_{\text{fusion}}$ are scalar hyperparameters balancing each loss term. This loss operates during training of LGST module in AeCSAD, updating the local and global student networks while keeping the teacher fixed.

### G. ANOMALY SCORE COMPUTATION AND NORMALIZATION

To evaluate anomalies at the image level, AeCSAD adopts a score fusion and normalization strategy aligned with the CSAD baseline. Both the LGST and Patch Histogram modules independently produce raw anomaly scores for each test image. For each module, validation scores are first used to compute normalization statistics. Specifically, trimmed $z$-score normalization is applied where the top and bottom 20% of validation scores are excluded to suppress outlier influence, and the remaining values are used to compute the mean $\mu_S$ and standard deviation $\sigma_S$. The normalized score $\hat{S}$ is then computed using

$$\hat{S} = \frac{S - \mu_S}{\sigma_S}. \quad (9)$$

This normalization is applied separately to each module's scores. The final anomaly score for each test image is then computed via additive fusion using

$$S_{\text{final}} = \hat{S}_{\text{hist}} + \hat{S}_{\text{LGST}}. \quad (10)$$

This scoring process is conducted independently for logical and structural anomaly types. To evaluate detection performance, binary ground-truth labels are defined based on the semantic annotation of each image, indicating whether it contains a logical anomaly, a structural anomaly, or is normal. The Area Under the Receiver Operating Characteristic curve (AUROC) is used as the primary evaluation metric, computed separately for each anomaly type and object category. A detailed discussion of performance using AUROC metric is presented in Section IV.

## IV. EXPERIMENTS

All experiments were conducted on a workstation with an NVIDIA Tesla V100 (32GB), Intel Xeon Silver 4114 CPU, Ubuntu 22.04, CUDA 12.5, and PyTorch 2.4.1. The environment was managed via Conda, with fixed random seeds across NumPy, PyTorch, and Python to ensure reproducibility. AeCSAD is evaluated on the MVTec LOCO AD benchmark [4], using resized $256\times256$ inputs normalized with ImageNet statistics, consistent with the pretrained WideResNet-50 backbone. We selected MVTec LOCO AD as the primary benchmark because it explicitly incorporates logical constraints such as correct object arrangements, presence or absence of components, and relational consistency. This makes it uniquely suited for evaluating methods that aim to detect both localized structural defects and context-aware relational violations. Rotations and flipping augmentations were applied during pseudo-label generation. Following CSAD, the semantic pseudo-label masks were regenerated using the official pipeline together with the referenced SAM-HQ and GroundingDINO models. As this procedure involves proposal-based segmentation and clustering, slight variations in the masks are expected when re-executed, a point also noted in the CSAD release. In this work, all CSAD and AeCSAD results are trained and evaluated using these same regenerated pseudo-labels to maintain a consistent and reproducible setting. The local and global student networks were trained using Adam optimizer with an initial learning rate of $1 \times 10^{-4}$, weight decay of $1 \times 10^{-4}$, and cosine annealing down to $1 \times 10^{-6}$. The same configuration was used for training the semantic segmentation model on pseudo-labeled masks. All models were trained with a batch size of 16 and a step decay schedule. Feature embeddings are aligned to $512 \times 64 \times 64$. The DeepLabV3+ segmentation network is trained in two stages over 120 epochs: the first 40 epochs use only $\mathcal{D}_L$ with LSA augmentations [20] and after epoch 40, $\mathcal{D}_U$ is incorporated. We evaluate AeCSAD using AUROC and computed it for each category in the MVTec LOCO AD dataset [4], reporting mean scores separately for logical and structural anomaly subsets. Our results are benchmarked against state-of-the-art methods including SimpleNet [32], PatchCore [13], AST [35], EfficientAD [38], ComAD [18], and CSAD. Comparative results are summarized in Table 1.

### A. RESULTS AND DISCUSSION

AeCSAD achieves a leading average AUROC of **95.05%** across both logical and structural anomaly categories, outperforming segmentation-based and segmentation-free baselines. In the logical anomaly category, it reaches an average AUROC of **97.56%**, with perfect or near-perfect scores on **Pushpins** (100.0%) and **Screw Bag** (99.7%), where anomalies involve misplaced or duplicated parts. This reflects its ability to model part co-occurrence and spatial configuration consistency via self-attention, surpassing segmentation-free baselines like SimpleNet and AST. It also performs strongly on **Splicing Connectors** (97.0%), where orientation or mirroring errors require relational reasoning. Notably, AeCSAD remains competitive in local anomaly categories such as **Juice Bottle** (95.5%) and **Breakfast Box** (95.6%), indicating that early convolutional layers preserve fine-grained sensitivity. Table 1 shows comparison of MVTec LOCO AD performance with state-of-the-art methods, as measured by image AUROC (%).

In the structural anomaly category, AeCSAD achieves an average AUROC of **92.54%**, ranking first among segmentation-based methods and second overall. On categories like **Screw Bag** (94.5%) and **Pushpins** (94.0%), it outperforms CSAD and ComAD, demonstrating its capacity to integrate local and global reasoning. While texture-specialized methods like EfficientAD slightly outperform AeCSAD on highly localized categories such as **Juice Bottle** (96.7%) and **Splicing Connectors** (91.3%), AeCSAD remains robust. In **Breakfast Box** (86.2%), where anomalies blend visual and semantic cues, AeCSAD still leads most baselines, highlighting its versatility across structural and logical anomaly types.

### B. QUALITATIVE RESULTS

Figures 7 and 8 show heatmaps over MVTec LOCO AD samples, highlighting AeCSAD's capacity to detect both logical and structural anomalies. In logical categories such as **Pushpins**, **Screw Bag**, and **Splicing Connectors**, global attention focuses precisely on duplicated, misaligned, or mirrored components, producing sharp red activations with minimal background noise, providing evidence of strong relational reasoning. In contrast, broader yellow spread in **Breakfast Box** suggests weaker semantic alignment, consistent with subtler layout errors.

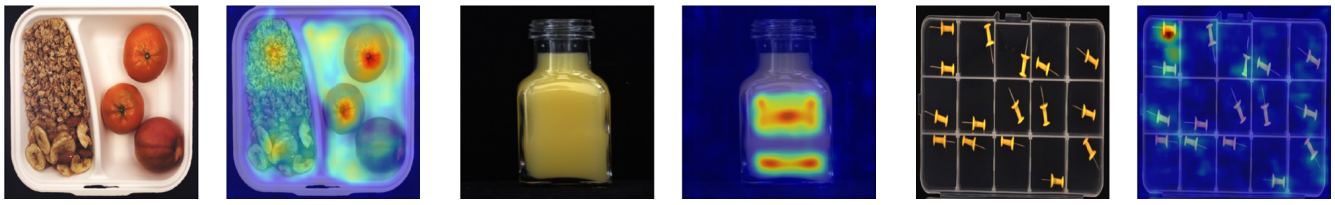
Structural anomalies are captured by the local student using pixel-aligned features and small receptive fields. As seen in **Pushpins** and **Screw Bag**, AeCSAD balances texture-level precision with spatial coherence, aiding detection in cluttered scenes. While segmentation-free models excel in simpler, texture-driven categories like **Juice Bottle** or **Splicing Connectors**, AeCSAD remains competitive. In **Breakfast Box**, where both reasoning modes are needed, the model preserves robust performance.

### C. ABLATION STUDIES

Table 2 summarizes the individual contributions of PatchHist, the LGST module in CSAD, and the self-attention enhanced LGST module in AeCSAD. PatchHist alone provides strong performance for logical anomalies (91.84% mean AUROC), as its histogram representation captures component-level frequency. However, it is less effective for structural anomalies (75.66%), where localized appearance consistency is essential. The LGST module from CSAD addresses this limitation by providing strong structural discrimination (91.33% mean AUROC), particularly in categories such as **Juice Bottle** (98.35%) and **Splicing Connectors** (96.05%).

**TABLE 1.** Comparison of MVTec LOCO AD performance with state-of-the-art methods, as measured by image AUROC (%). The "AeCSAD" column presents the results of our method. Bold value represents top value across categories or groups. A value underlined represents second to top across categories or groups.

| | No Segmentation | | | | Segmentation | | |
|---|---|---|---|---|---|---|---|
| **Category** | SimpleNet | PatchCore | AST | EfficientAD | ComAD | CSAD | AeCSAD (Ours) |
| **Logical Anomalies (LA)** | | | | | | | |
| Breakfast Box | 77.1 | 74.8 | 80.0 | 85.5 | 91.1 | **95.8** | 95.6 |
| Juice Bottle | 87.8 | 93.9 | 91.6 | **98.4** | 95.0 | 95.4 | 95.5 |
| Pushpins | 69.0 | 63.6 | 65.1 | 97.7 | 95.7 | 99.6 | **100.0** |
| Screw Bag | 51.6 | 57.8 | 80.1 | 56.7 | 71.9 | 99.3 | **99.7** |
| Splicing Connectors | 72.0 | 79.2 | 81.8 | *95.5* | 93.3 | 95.7 | **97.0** |
| **Average (Logical)** | 71.5 | 73.9 | 79.7 | 86.8 | 89.4 | 97.16 | **97.56** |
| **Structural Anomalies (SA)** | | | | | | | |
| Breakfast Box | 80.9 | 80.1 | 79.9 | **88.4** | 81.6 | 85.9 | 86.2 |
| Juice Bottle | 90.4 | 98.5 | 95.5 | **99.7** | 98.2 | 97.3 | 96.7 |
| Pushpins | 81.6 | 87.9 | 77.8 | **96.1** | 91.1 | 93.2 | 94.0 |
| Screw Bag | 83.3 | 92.0 | **95.9** | 90.7 | 88.5 | 91.8 | 94.5 |
| Splicing Connectors | 82.6 | 88.0 | 89.4 | **98.5** | 94.9 | 92.4 | 91.3 |
| **Average (Structural)** | 83.7 | 89.3 | 87.7 | **94.7** | 90.9 | 92.12 | 92.54 |
| **Total Average** | 77.6 | 81.6 | 83.7 | 90.7 | 90.1 | 94.64 | **95.05** |



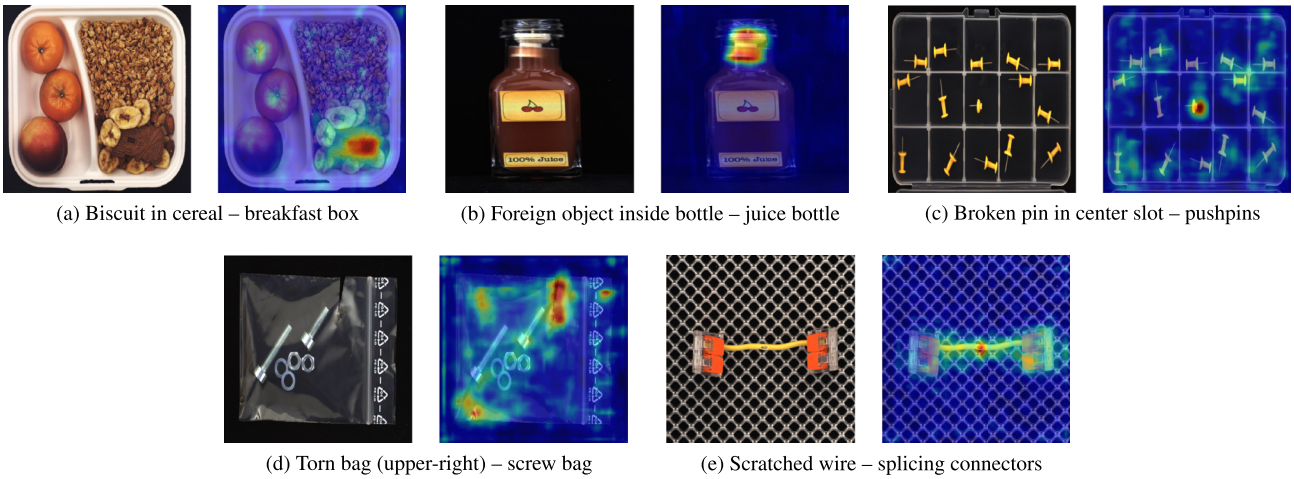
(a) Misplaced fruit – breakfast box (b) Missing label – juice bottle (c) Two pins in one compartment – pushpins

(d) Duplicate large screws – screw bag (e) (e) Miswired cable – splicing connectors

**FIGURE 7.** Logical anomalies detected by AeCSAD. Each pair shows the original image (left) and its attention-based anomaly map (right). AeCSAD highlights semantic inconsistencies such as missing parts, duplications, or miswirings across categories in MVTec LOCO AD.

However, because its global student is constrained by a CNN autoencoder, it remains less effective on logical anomaly categories whose irregularities arise from spatial configuration and part-to-part relationships, mostly evident in **Pushpins**.

AeCSAD strengthens this component by introducing self-attention into the global student. For logical anomalies, the benefit is most visible in categories where the anomaly mechanism is relational rather than appearance-based. In **Pushpins**, for example, logical AUROC improves



(a) Biscuit in cereal – breakfast box (b) Foreign object inside bottle – juice bottle (c) Broken pin in center slot – pushpins

(d) Torn bag (upper-right) – screw bag (e) Scratched wire – splicing connectors

**FIGURE 8.** Structural anomalies detected by AeCSAD. Each pair shows the original image (left) and its attention-based anomaly map (right). AeCSAD detects breakages, foreign objects, scratches, and packaging defects with high precision.

**TABLE 2.** Ablation comparison of PatchHist, LGST (CSAD), and LGST (AeCSAD). We report image-level AUROC (%) for logical (LA) and structural (SA) anomalies on MVTec LOCO AD. PatchHist outputs are identical across LGST variants.

| Category | PatchHist | LGST (CSAD) | LGST (AeCSAD) |
|---|---|---|---|
| **Logical Anomalies (LA)** | | | |
| Breakfast Box | **95.62** | 84.49 | 86.64 |
| Juice Bottle | 87.92 | **95.81** | 95.63 |
| Pushpins | 90.63 | 98.89 | **98.98** |
| Screw Bag | **99.79** | 66.23 | 59.57 |
| Splicing Connectors | 85.27 | 95.36 | **97.25** |
| **Mean** | **91.84** | 88.16 | 87.61 |
| **Structural Anomalies (SA)** | | | |
| Breakfast Box | 75.44 | 86.50 | **86.91** |
| Juice Bottle | 75.34 | **98.35** | 98.04 |
| Pushpins | 71.72 | 92.92 | **94.02** |
| Screw Bag | 85.52 | 82.81 | **88.89** |
| Splicing Connectors | 70.30 | **96.05** | 93.84 |
| **Mean** | 75.66 | 91.33 | **92.34** |

from 98.89% to 98.98%, and **Splicing Connectors** shows a similar improvement. These gains reflect the ability of self-attention to model interactions among spatially distant components. For structural anomalies, LGST (AeCSAD) achieves the highest mean AUROC (92.34%), driven by its ability to reconstruct global appearance patterns rather than by relational effects.

Self-attention does not provide uniform improvements across all settings. For example, in **Breakfast Box** and **Screw Bag**, performance gains are modest or neutral, and the logical AUROC in **Screw Bag** decreases. These outcomes are consistent with the fact that PatchHist already captures high-level irregularities in such categories, and additional relational modeling yields limited benefit.

Beyond image-level AUROC, we also assess how self-attention influences the spatial coherence of anomaly localization using the Per-Region Overlap (PRO) metric.

**TABLE 3.** PRO-based localization performance on MVTec LOCO AD for the LGST module. Values are reported as (logical PRO / structural PRO).

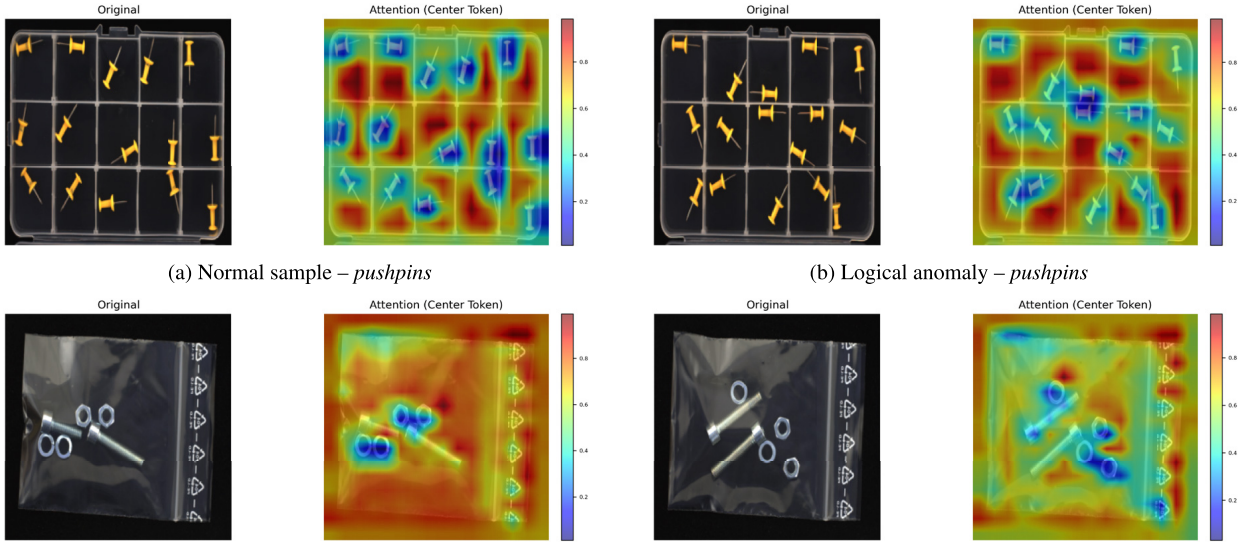
| Category | LGST (CSAD) | LGST (AeCSAD) |
|---|---|---|
| Breakfast Box | 68.9 / 85.8 | 71.0 / 86.4 |
| Juice Bottle | 80.0 / 93.3 | 80.0 / 93.1 |
| Pushpins | 75.0 / 75.4 | 77.4 / 76.3 |
| Screw Bag | 24.3 / 72.3 | 50.7 / 89.9 |
| Splicing Connectors | 60.5 / 91.7 | 59.8 / 91.5 |
| **Mean** | 61.7 / 83.7 | **67.8 / 87.4** |

Table 3 reports logical and structural PRO scores for LGST (CSAD) and LGST (AeCSAD). Consistent with the AUROC trends, the self-attention variant achieves higher mean PRO for both logical (67.8% vs. 61.7%) and structural anomalies (87.4% vs. 83.7%). The largest gains occur in categories with distributed or relational structure, such as *screw bag* and *pushpins*, further indicating that long-range dependency modeling benefits anomaly mechanisms that involve interactions among multiple components.

To further contextualize these findings, Figure 9 presents representative qualitative attention responses from the self-attention global student in AeCSAD. In normal samples, center-token attention distributes smoothly across components that form coherent spatial and relational patterns. In logical anomaly samples, attention shifts toward duplicated, missing, or misconfigured elements, reflecting the relational inconsistencies that define these categories. These patterns mirror the PRO gains and show that self-attention strengthens the reconstruction dynamics of LGST, especially for relational anomalies.

## V. LIMITATIONS

AeCSAD relies on foundation models, such as SAM and GroundingDINO, which are susceptible to failures under




(a) Normal sample – *pushpins*

(b) Logical anomaly – *pushpins*

(c) Normal sample – *screw_bag*

(d) Logical anomaly – *screw_bag*

**FIGURE 9. Qualitative effect of self-attention within the LGST branch. Each pair shows the original image with its center-token attention overlay. In normal samples, attention distributes coherently across relationally consistent components. In logical anomalies, attention shifts toward duplicated, missing, or structurally inconsistent elements. These patterns complement the PRO improvements observed in Table 3, illustrating how self-attention enhances the relational coherence of reconstructed features.**

occlusion, clutter, or poor lighting. These failures can introduce noise in the early segmentation pipeline, and despite filtering, such noise may propagate and affect downstream anomaly detection. Moreover, the Patch Histogram module lacks rotational and positional invariance and is sensitive to bin granularity. This reduces its robustness in settings with natural part variability, such as deformable objects or multi-view inspections. In addition, the self-attention mechanism requires sufficient training diversity to model spatial-semantic relationships effectively. Any rare configurations may be incorrectly flagged due to distributional unfamiliarity. AeCSAD employs fixed fusion weights based on the assumption of equal contribution from each module regardless of the anomaly type. This design choice may reduce the model's adaptability in scenarios where anomaly instances are predominantly logical or structural in nature.

Another limitation stems from the assumption of RGB-formatted images. This design choice is primarily dictated by the structure of the MVTec LOCO AD dataset, which provides all samples in RGB format. The RGB modality captures perceptual cues such as color, shading, and texture, which are often useful for accurate image segmentation and for modeling inter-object relationships. These visual features support the identification of semantically inconsistent object configurations, including misplaced, missing, or contextually inappropriate items. In scenarios where grayscale images are employed instead of RGB, such perceptual cues are substantially reduced or eliminated. Consequently, both anomaly detection accuracy and the model's capacity for relational reasoning are adversely impacted. AeCSAD also lacks an explicit mechanism for uncertainty estimation. While anomaly scores approximate deviation severity, they do not ensure recognition of failure cases.

## VI. CONCLUSION

This work presented AeCSAD, an anomaly detection framework that extends CSAD [20] by incorporating self-attention into the global student network. This modification enables long-range contextual reasoning while preserving spatial granularity, allowing the model to detect anomalies that violate semantic relationships. AeCSAD adopts a modular design that fuses independently normalized scores from the Patch Histogram module and the enhanced Local-Global Student-Teacher (LGST) architecture. Each component targets a distinct anomaly modality: the histogram module captures distributional shifts, the local student models fine-grained texture defects, and the self-attention-enhanced global student detects layout and relational inconsistencies. This decoupled structure leads to improved AUROC scores across both logical and structural anomaly categories on the MVTec LOCO AD benchmark. The key insight is that architectural modularity combined with explicit spatially aware feature learning and relational reasoning yields better visual anomaly detection.

## FUTURE WORK

While AeCSAD demonstrates strong performance, several avenues remain for improvement. Future research could explore confidence-aware filtering strategies, such as intra-cluster consistency or cross-model agreement, to reduce supervision noise. Although self-attention strengthens distant spatial reasoning, it lacks explicit part-level representation; incorporating structured models such as slot attention [44]

or component-wise graph neural networks [45] may enable more compositional and interpretable forms of logical reasoning. Beyond the category-specific setting used in this work, extending AeCSAD to few-shot approaches [46] or contrastive adaptation [47] remains a promising direction. The static use of fusion weights across branches could also be revisited, with dynamic weighting schemes that adapt to the anomaly type or context offering greater robustness. Another important direction is supporting grayscale imaging and confidence-aware learning to increase reliability in safety-critical industrial deployments. In addition, replacing ImageNet pretraining with domain-specific self-supervised segmentation [48] could improve part-level feature quality and enhance robustness under industrial variation.

## REFERENCES


[1] M. Muro, R. Maxim, and J. Whiton. (2019). *America's Advanced Industries: New Trends*. [Online]. Available: https://www.brookings.edu/research/americas-advanced-industries-new-trends

[2] E. A. Lee, "Cyber physical systems: Design challenges," in *Proc. 11th IEEE Int. Symp. Object/Compon./Service-Oriented Real-Time Distrib. Comput. (ISORC)*, Jul. 2008, pp. 363–369.

[3] H. Jung, S. Yoon, S. Lee, and J. Kim, "Industrial anomaly detection: Benchmark dataset and method," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2021, pp. 4483–4492.

[4] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, and C. Steger, "Beyond dents and scratches: Logical constraints in unsupervised anomaly detection and localization," *Int. J. Comput. Vis.*, vol. 130, no. 4, pp. 947–969, Apr. 2022.

[5] G. Pang, C. Shen, L. Cao, and A. van den Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surveys (CSUR)*, vol. 54, no. 2, pp. 1–38, 2021.

[6] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educ. Psychol.*, vol. 24, no. 6, pp. 417–441, Sep. 1933.

[7] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. Ser. B, Stat. Methodol.*, vol. 39, no. 1, pp. 1–22, Sep. 1977.

[9] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, Dec. 1989.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[11] W. Luo, Y. Li, R. Urtasun, and R. S. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2017, pp. 4898–4906.

[12] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad—A comprehensive real-world dataset for unsupervised anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9592–9600.

[13] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler, "Towards total recall in industrial anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14298–14308.

[14] D. Gudovskiy, S. Ishizaka, and K. Kozuka, "CFLOW-AD: Real-time unsupervised anomaly detection with localization via conditional normalizing flows," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 1819–1828.

[15] T. Defard, A. Setkov, A. Loesch, and R. Audigier, "PaDiM: A patch distribution modeling framework for anomaly detection and localization," in *Proc. Int. Conf. Pattern Recognit. (ICPR) Workshops*, 2021, pp. 475–489.

[16] J. Yu, Y. Zheng, X. Wang, W. Li, Y. Wu, R. Zhao, and L. Wu, "FastFlow: Unsupervised anomaly detection and localization via 2D normalizing flows," in *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2021, pp. 2813–2822.

[17] X. Tao, X. Gong, X. Zhang, S. Yan, and C. Adak, "Deep learning for unsupervised anomaly localization in industrial images: A survey," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–21, 2022.

[18] T. Liu, B. Li, X. Du, B. Jiang, X. Jin, L. Jin, and Z. Zhao, "Component-aware anomaly detection framework for adjustable and logical industrial visual inspection," *Adv. Eng. Informat.*, vol. 58, Oct. 2023, Art. no. 102161.

[19] S. Kim, S. An, P. Chikontwe, M. Kang, E. Adeli, K. M. Pohl, and S. H. Park, "Few shot part segmentation reveals compositional logic for industrial anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 8591–8599.

[20] Y.-H. Hsieh and S. Lai, "CSAD: Unsupervised component segmentation for logical anomaly detection," in *Proc. 35th Brit. Mach. Vis. Conf. (BMVC)*, 2024, pp. 1–14.

[21] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu, Y. Guo, and L. Zhang, "Recognize anything: A strong image tagging model," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 1724–1732.

[22] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Marrying dino with grounded pre-training for open-set object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, 2024, pp. 38–55.

[23] A. M. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W. Lo, P. Dollár, and R. Girshick, "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 3992–4003.

[24] C. Zhou and R. C. Paffenroth, "Anomaly detection with robust deep autoencoders," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2017, pp. 665–674.

[25] H. Zhao, J. Jia, and V. Koltun, "Exploring self-attention for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10073–10082.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[27] N. Cohen and Y. Hoshen, "Sub-image anomaly detection with deep pyramid correspondences," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4813–4822.

[28] C. Huang, H. Guan, A. Jiang, Y. Zhang, M. Spratling, and Y. Wang, "Registration based few-shot anomaly detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 303–319.

[29] J. Bae, J.-H. Lee, and S. Kim, "PNI: Industrial anomaly detection using position and neighborhood information," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*. Chicago, IL, USA: Industrial, Oct. 2023, pp. 6350–6360. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/BaePNIIndustrialAnomalyDetectionusing PositionandNeighborhoodInformationICCV2023paper.html

[30] J. Hyun, S. Kim, G. Jeon, S. H. Kim, K. Bae, and B. J. Kang, "Reconpatch: Contrastive patch representation learning for industrial anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Aug. 2024, pp. 2052–2061.

[31] D. McIntosh and A. B. Albu, "Inter-realization channels: Unsupervised anomaly detection beyond one-class classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6262–6272. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/McIntoshInter-RealizationChannelsUnsupervisedAnomalyDetection BeyondOne-ClassClassificationICCV2023paper.html

[32] Z. Liu, Y. Zhou, Y. Xu, and Z. Wang, "SimpleNet: A simple network for image anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 20402–20411.

[33] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student–teacher anomaly detection with discriminative latent embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4183–4192.

[34] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9727–9736.

[35] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Asymmetric student–teacher networks for industrial anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Oct. 2023, pp. 2591–2601.
[36] Z. Gu, L. Liu, X. Chen, R. Yi, J. Zhang, Y. Wang, C. Wang, A. Shu, G. Jiang, and L. Ma, "Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16355–16363. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2023/html/GuRememberingNormalityMemory-guidedKnowledgeDistillationforUnsupervisedAnomalyDetectionICCV2023paper.html
[37] S. Wang, L. Wu, L. Cui, and Y. Shen, "Glancing at the patch: Anomaly localization with global and local feature comparison," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 254–263. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/WangGlancingatthePatchAnomalyLocalizationWithGlobalandLocalCVPR2021paper.html
[38] K. Batzner, L. Heckler, and R. König, "EfficientAD: Accurate visual anomaly detection at millisecond-level latencies," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 127–137.
[39] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790–799, Aug. 1995.
[40] L. McInnes, J. Healy, and S. Astels, "Hdbscan: Hierarchical density based clustering," *J. Open Source Softw.*, vol. 2, no. 11, p. 205, Mar. 2017, doi: 10.21105/joss.00205.
[41] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 87.1–87.12.
[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[43] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 833–851.
[44] F. Locatello, D. Weissenborn, T. Unterthiner, A. Mahendran, G. Heigold, J. Uszkoreit, A. Dosovitskiy, and T. Kipf, "Object-centric learning with slot attention," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020, pp. 11519–11530. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/8511df98c02ab60aea1b2356c013bc0f-Abstract.html
[45] Q. Zhang, R. Cao, Y. Wu, and S. Zhu, "Growing interpretable part graphs on ConvNets via multi-shot learning," in *Proc. Thirty-First AAAI Conf. Artif. Intell. (AAAI-17)*, vol. 31, 2017, pp. 2898–2906.
[46] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 1126–1135.
[47] D. Peng, Z. Zhang, P. Hu, Q. Ke, D. K. Y. Yau, and J. Liu, "Harnessing text-to-image diffusion models for category-agnostic pose estimation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 342–360.
[48] A. Ziegler and Y. M. Asano, "Self-supervised learning of object parts for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14482–14491.

**HAFIZ A. AMJAD** received the B.S. degree in computer science from Bahauddin Zakariya University (BZU), Multan, Pakistan, in 2018, the M.S. degree (Hons.) in computer science from the National University of Computer and Emerging Sciences (NUCES), Lahore, Pakistan, in 2020, and the second M.S. degree in computer science from Ontario Tech University, Canada, in 2025. He has experience in developing and integrating AI agents into VR environments and has served as a Lecturer of the course theory of computation at NUCES, where he was nominated as the Best Lecturer. His current research interests include computer vision and natural language processing.

**IRIDA SHALLARI** (Member, IEEE) received the joint M.Sc. degree in embedded systems from the University of Southampton, U.K., and NTNU, Norway, in 2016, and the Ph.D. degree in embedded vision systems from Mid Sweden University, Sweden. In 2023 and 2024, she was a Visiting Researcher at Ontario Tech University and the University of Salerno, investigating anomaly detection and AI-based measurement systems. She is currently an Assistant Professor with Mid Sweden University. She has co-authored around 30 publications. She has garnered competitive research funding through both national and international grants. Her research interests include the design and optimization of AI-based IoT nodes intended for intelligent monitoring, resource-constrained computations, and robust fault detection, often in challenging domains (e.g., railroad condition monitoring and snow depth measurement systems).

**MATTIAS O' NILS** (Member, IEEE) received the B.S. degree in electrical engineering from Mid Sweden University, Sundsvall, Sweden, in 1993, and the Licentiate and Ph.D. degrees in electronic systems design from the Royal Institute of Technology, Stockholm, Sweden, in 1996 and 1999, respectively. He is currently a Professor with the Department of Computer and Electrical Engineering and leads a research group in AI-based IoT systems at Mid Sweden University. His current research interests include design methods and implementation of embedded DNN-based systems, especially in the implementation of real-time video processing and time series processing systems.

**FAISAL Z. QURESHI** (Senior Member, IEEE) received the B.Sc. degree in mathematics and physics from Punjab University, Lahore, Pakistan, in 1993, the M.Sc. degree in electronics from Quaid-e-Azam University, Islamabad, Pakistan, in 1995, and the M.Sc. and Ph.D. degrees in computer science from the University of Toronto, Toronto, Canada, in 2000 and 2007, respectively. He is currently a Professor of computer science with the Faculty of Science, Ontario Tech University, where he leads the Visual Computing Laboratory. His scientific and engineering interests center on the study of computational models of visual perception to support autonomous, purposeful behaviour in the context of ad hoc networks of smart cameras. His research interest includes computer vision. He is a member of ACM and the Secretary and a member of CIPPRS. He is also active in journal special issues and conference organizations. He served as the General Co-Chair for the Workshop on Camera Networks and Wide-Area Scene Analysis (co-located with CVPR), from 2011 to 2013. He also served as the Co-Chair of the Computer and Robot Vision (CRV) Conference 2015/16 Meetings.