GeoMAD: Geometry-Aware Multi-View Anomaly Detection via Deformable Fusion and Distributional Alignment

Shang-Fu Chen[1]
chenshangfu@cmlab.csie.ntu.edu.tw
Jhih-Ciang Wu[2]
jcwu@csie.ntnu.edu.tw
Kuan-Chuan Peng[3]
kpeng@merl.com
Wen-Huang Cheng[1,5]
wenhuang@csie.ntu.edu.tw
Kai-Lung Hua[4,6]
kai.hua@microsoft.com

[1] National Taiwan University
Taipei, Taiwan
[2] National Taiwan Normal University
Taipei, Taiwan
[3] Mitsubishi Electric Research Laboratories (MERL)
Cambridge, MA, USA
[4] Microsoft Taiwan Corporation
Taipei, Taiwan
[5] VinUniversity
Hanoi, Vietnam
[6] National Taiwan University of Science and Technology
Taipei, Taiwan

Abstract

Multi-view anomaly detection (MvAD) detects defects by exploiting complementary observations from multiple camera viewpoints. The central challenge is to fuse views with sufficient geometric awareness while remaining scalable to multi-class industrial settings. Existing methods typically fall into two extremes: voxel-based fusion provides explicit geometric alignment but requires costly 3D construction and class-specific assumptions, whereas lightweight patch-based fusion is efficient but relies on discrete candidate matching and lacks continuous cross-view correspondence. In this paper, we propose `GeoMAD`, a unified multi-view, multi-class AD framework that addresses both geometric correspondence deficiency and distributional inconsistency. Our *Cross-view Deformable Fusion Module* (CDFM) learns content-adaptive, view-pair-specific sampling offsets directly on 2D feature maps and arranges them across a multi-scale window pyramid with image-global reference sampling, enabling hierarchical cross-view correspondence without camera calibration, voxel construction, or class-specific 3D supervision. We further introduce *Distributional View Alignment* (DVA), a self-supervised cross-view regularization loss that aligns each view's bottleneck distribution against a per-instance view-centric target, enforcing global consistency without pixel-level correspondence. Together, CDFM and DVA bridge local geometric correspondence and global distributional consistency, providing geometry-aware and distribution-consistent fusion while preserving the efficiency of 2D feature-space learning. Extensive experiments on Real-

© 2026. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

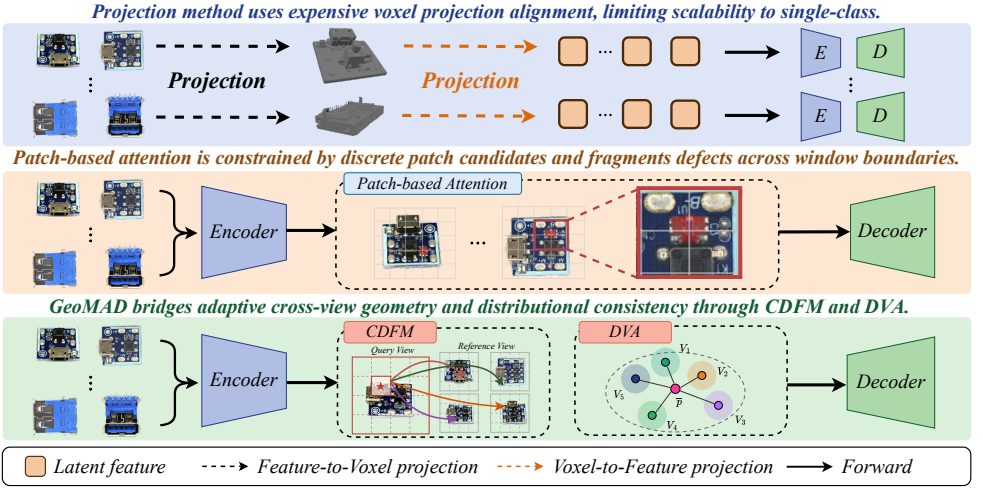

Figure 1: GeoMAD combines Cross-view Deformable Fusion Module (CDFM) and Distributional View Alignment (DVA) to achieve geometry-aware and distribution-consistent fusion directly in 2D feature space, addressing key limitations of existing methods [11, 20].

IAD and MANTA-Tiny show that GeoMAD achieves strong detection and localization performance in unified MvAD.

1 Introduction

Anomaly Detection (AD) is a fundamental task in industrial inspection [14, 21, 23, 33], medical imaging [2, 3, 35], and video surveillance [13, 15, 19, 38], where the goal is to identify deviations from normal patterns before they lead to costly failures or safety risks. Since abnormal samples are rare and diverse, most AD methods [1, 26, 39] are trained using only normal data and detect anomalies by measuring deviations from learned normality. Existing one-class methods [4, 16, 17, 22, 25, 27, 36] often train a separate model for each object class. Although effective, this design becomes expensive to deploy across multiple product lines. Recent unified AD methods [8, 29, 34] address this issue by training a single model across multiple classes, but they still largely operate under a single-view assumption.

Single-view imagery is inherently limited for anomaly detection. Defects may be occluded, weakly illuminated, or geometrically ambiguous from one camera angle, while being clearly visible from another. Multi-view anomaly detection (MvAD)[1] addresses this limitation by leveraging multiple synchronized observations of the same object. Given views $\{X_1, \ldots, X_V\}$, MvAD aims to exploit their complementary information for more reliable anomaly detection and localization than any single view. Recent multi-view datasets such as Real-IAD [28] and MANTA [7] have made this setting increasingly practical. However, effective MvAD requires more than simply processing multiple images. The key challenge is how to fuse views in a way that is both geometrically meaningful and computationally scalable. Existing cross-view fusion strategies expose a clear trade-off between geometric fidelity and computational scalability, as illustrated in Figure 1. Voxel-projection methods [20] establish explicit cross-view correspondence by lifting features into a shared 3D representation. This provides strong

[1]We use MvAD to denote the Multi-view Anomaly Detection task, while MVAD refers to a previous work [11].

geometric alignment, but requires expensive voxel construction and is often tied to class-specific 3D assumptions. Patch-attention methods [11], by contrast, are lightweight and more general, but their correspondence is still mediated by discrete patch-level candidate selection. As a result, they lack an adaptive mechanism for cross-view geometry and may fragment anomalies that cross patch or window boundaries. These limitations are particularly problematic for MvAD, where the same surface defect can appear at different image locations, scales, and shapes across views.

We identify two issues that an effective multi-view AD model should address. The first is *geometric correspondence deficiency*: lightweight patch-based fusion can retrieve cross-view candidates through patch-level similarity and then refine them with pixel-wise attention, but the correspondence remains constrained by discrete patch candidates rather than continuous geometric displacement. Such coarse candidate selection provides limited adaptability to viewpoint-dependent changes in location, scale, and shape, and may still fragment a continuous defect when its corresponding evidence is distributed across multiple patch regions. The second is *distributional inconsistency*: even after local cross-view information is exchanged, the fused representations from different views are not explicitly encouraged to share compatible normality statistics. Pixel-level consistency is not suitable for resolving this issue, because the same physical surface region may appear at different spatial locations across cameras. Instead, multi-view AD requires both adaptive local geometric correspondence and global distributional consistency.

To address these challenges, we propose `GeoMAD`, a unified multi-view, multi-class AD framework built upon reverse distillation. `GeoMAD` introduces *Cross-view Deformable Fusion Module* (CDFM), a bottleneck design that learns content-adaptive, view-pair-specific sampling offsets with image-global reference reach across a hierarchical window pyramid. By sampling reference-view features from continuous, content-adaptive locations at multiple scales, CDFM provides implicit geometric reasoning without camera calibration, voxel construction, or class-specific 3D supervision. We further introduce *Distributional View Alignment* (DVA), a self-supervised cross-view regularization loss that aligns each view's bottleneck distribution against a per-instance view-centric target. The synchronized views of the same normal object provide a natural supervisory signal, allowing DVA to encourage compatible feature statistics without anomaly labels, pixel-level matching, or explicit 3D projection. CDFM learns local deformable correspondence, while DVA enforces global view-level consistency. Together, they enable geometry-aware and distribution-consistent multi-view AD without the memory and supervision requirements of explicit 3D alignment. We evaluate `GeoMAD` on Real-IAD [28] and MANTA-Tiny [7]. Extensive experiments and ablation studies show that `GeoMAD` achieves strong anomaly detection and localization performance in unified multi-view AD, while maintaining practical computational efficiency.

Our main contributions are summarized as follows:

- We propose `GeoMAD`, a unified multi-view, multi-class AD framework that jointly addresses adaptive cross-view correspondence and view-level distributional consistency without explicit 3D reconstruction.
- We propose *Cross-view Deformable Fusion Module* (CDFM), a calibration-free 2D bottleneck design that learns content-adaptive, view-pair-specific sampling offsets with image-global reach for geometry-aware multi-view fusion.
- We introduce *Distributional View Alignment* (DVA), which regularizes cross-view consistency in distribution space via a per-instance view-centric target, without requiring pixel-level correspondence or explicit 3D projection.

Method Properties	AD Categories				
	$\mathcal{G}_0$	$\mathcal{G}_1$	IDIF	MVAD	GeoMAD (ours)
Single model for all classes	✗	✓	✗	✓	✓
Designed for multi-view AD	✗	✗	✓	✓	✓
Learned 2D geometric correspondence	✗	✗	✗	✗	✓
Distributional view alignment	✗	✗	✗	✗	✓

Table 1: Characteristic comparison between GeoMAD (ours) and prior AD works. $\mathcal{G}_0$ denotes class-specific single-view AD methods [4, 5, 6, 12, 17, 24, 32]; $\mathcal{G}_1$ denotes unified single-view AD methods [8, 9, 10, 29, 31, 34]. IDIF [20] and MVAD [11] denote representative multi-view AD methods.

- Extensive experiments on Real-IAD and MANTA-Tiny demonstrate that GeoMAD improves unified multi-view anomaly detection and localization over existing methods.

2 Related Work

Unsupervised anomaly detection: Unsupervised AD aims to identify anomalies and localize defects using only normal samples during training, assuming anomalies are rare and can be distinguished from learned normal patterns. Existing AD methods can be broadly grouped into class-specific and unified settings, as summarized in Table 1. Class-specific approaches $\mathcal{G}_0$ train separate models for individual object categories, achieving strong performance but incurring high storage and deployment costs. Unified approaches $\mathcal{G}_1$ use a single model across categories, improving scalability while introducing potential inter-class interference. Category $\mathcal{G}_0$ includes works [4, 5, 6, 12, 17, 24, 32] that use separate models for each object class. These methods can be further organized into three styles, *i.e.* , synthesis-based, embedding-based, and reconstruction-based approaches. Synthesis-based methods [12, 17] create artificial anomalies during training but suffer from synthetic-to-real domain gaps and fail to capture cross-view spatial relationships. Embedding-based methods [5, 24] encode images into feature representations using pre-trained networks; however, they encounter feature misalignment issues in multi-view settings where features from different viewpoints lack spatial correspondence. Reconstruction-based methods [4, 6, 32] train encoder-decoder networks to reconstruct normal patterns, but may inadvertently reconstruct abnormal regions, leading to false negatives that are exacerbated in multi-view contexts without proper cross-view constraints. Category $\mathcal{G}_1$ encompasses recent works [8, 9, 10, 29, 31, 34] that require only a single model for all object classes. ViTAD [34] utilizes Vision Transformers for feature regression, while MambaAD [9] employs state space models for both global and local modeling. While these unsupervised methods have shown efficacy in unified scenarios [18], they typically struggle with multi-view designs, often resorting to single-view voting ensembles that process each view independently and yield suboptimal sample-level scores. Compared to these previous methods, we propose a unified architecture that carries out multi-view AD while explicitly exploring correlations across different viewpoints.

Multi-view anomaly detection: Building on single-view AD, recent works address the harder multi-view setting. MVAD [11] uses patch-based cross-attention for adaptive multi-view selection, but suffers from *geometric correspondence deficiency*: patch partitioning fragments boundary anomalies, causing incomplete detection. It also lacks cross-view constraints, yielding *distributional inconsistency* where representations from different views fail to share compatible normality statistics. IDIF [20] aligns views via voxel projection, but requires

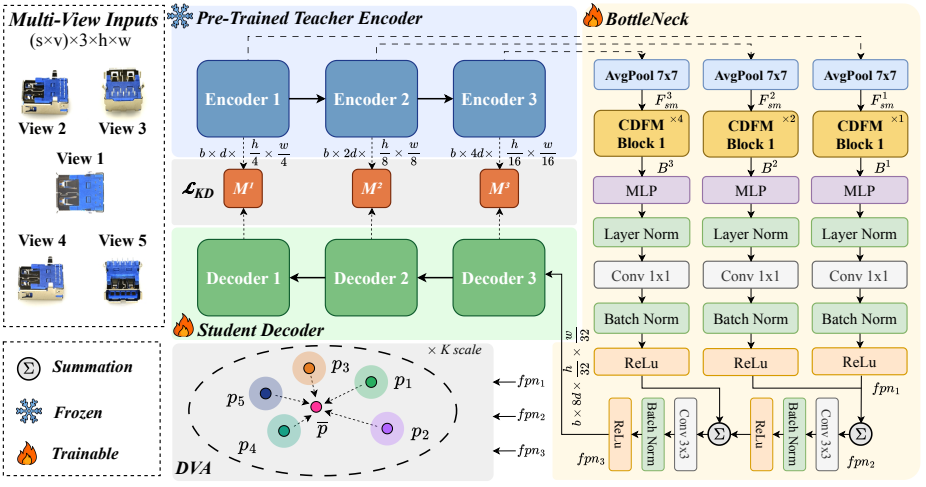

Figure 2: Overview of our proposed GeoMAD. Given multi-view input images, GeoMAD adopts a teacher-student architecture with the Cross-view Deformable Fusion Module (CDFM) for deformable cross-view feature fusion. Distributional View Alignment (DVA) mitigates distributional inconsistency by aligning bottleneck feature distributions across views. Multi-view anomaly scores are then aggregated for final detection.

costly voxel transformations and is limited to single-class settings, as diverse geometries hinder unified 3D modeling. As shown in Table 1, GeoMAD unifies learned 2D cross-view correspondence via CDFM and distributional view alignment via DVA, avoiding explicit 3D reconstruction while supporting multi-class AD.

3 Our proposed Method – GeoMAD

3.1 Overview

Let $\mathcal{D}_{\text{train}} = \{\mathcal{I}_1, \ldots, \mathcal{I}_n\}$ denote a set of n anomaly-free multi-view training instances, where each instance $\mathcal{I}_i = \{I_i^1, \ldots, I_i^V\}$ contains V views of the same object. Let $\mathcal{D}_{\text{test}} = \{\mathcal{I}_{n+1}, \ldots, \mathcal{I}_{n+m}\}$ denote a test set containing both normal and anomalous instances. The goal is to learn, using only anomaly-free training data, a model that can detect and localize anomalies in the test set. Following [11, 20], we formulate multi-view AD by organizing each sample as a set of synchronized views. Unlike single-view AD methods that process a batch of B independent images, multi-view methods restructure the batch as $B = S \times V$, where S is the number of object instances and V is the number of views per instance. This organization enables cross-view interaction within each object instance while preserving efficient batched computation. As illustrated in Figure 2, GeoMAD is designed for multi-view, multi-class AD. Following the reverse distillation paradigm of RD4AD [6], GeoMAD consists of three core components: a pretrained teacher network, a trainable student network, and a cross-view bottleneck module based on our proposed *Cross-view Deformable Fusion Module* (CDFM).

Given a multi-view input $\mathbf{X} \in \mathbb{R}^{B\times 3\times h\times w}$, GeoMAD processes each view through the teacher and student networks. The frozen teacher extracts multi-scale features $\mathbf{T}^k \in \mathbb{R}^{B\times d_k\times h_k\times w_k}$, where $k \in \{1, \ldots, K\}$ indexes the feature scale. The CDFM bottleneck reshapes these features

into multi-view form, $\mathbb{R}^{S\times V\times d_k\times h_k\times w_k}$, and performs cross-view fusion at each scale to produce fused bottleneck representations $\mathbf{B}^k$. These multi-scale representations are merged through a top-down feature-pyramid fusion into a single coarsest-scale feature, which conditions the trainable student decoder that reconstructs multi-scale features $\mathbf{S}^k \in \mathbb{R}^{B\times d_k\times h_k\times w_k}$ matching the teacher's scales. The training objective combines the multi-scale reverse-distillation loss $\mathcal{L}_{\text{KD}}$ with the proposed Distributional View Alignment loss $\mathcal{L}_{\text{DVA}}$ (Sec. 3.3), which regularizes view-level feature statistics within each multi-view instance. $\mathcal{L}_{KD}$ measures feature alignment between the teacher and student networks across K scales:

$$\mathcal{L}_{KD} = \sum_{k=1}^{K} \left\{ \frac{1}{H_k W_k} \sum_{h=1}^{H_k} \sum_{w=1}^{W_k} M^k(h,w) \right\}, \tag{1}$$

where $M^k(h,w)$ is the cosine distance between the teacher and student features at spatial location (h,w) and scale k. The multi-scale formulation captures anomalies at different semantic and spatial resolutions, while $\mathcal{L}_{\text{DVA}}$ encourages distributional consistency across views of the same object instance.

At inference, GeoMAD generates dense pixel-level anomaly scores $\mathbf{A}_{px} \in \mathbb{R}^{mV\times H\times W}$ by aggregating the multi-scale cosine distances $M^k(y,x)$ between teacher and student features. The image-level anomaly score $\mathbf{A}_{im} \in \mathbb{R}^{mV}$ is obtained by spatial max-pooling over each view. The final sample-level anomaly score $\mathbf{A}_{sa} \in \mathbb{R}^{m}$ is computed by first applying spatial max-pooling across each view, then averaging across all V viewpoints within each sample, effectively capturing the most abnormal perspective while leveraging multi-view consensus.

3.2 Cross-view Deformable Fusion Module

Multi-view AD requires cross-view interaction that is both geometrically meaningful and computationally efficient. Existing fusion strategies usually satisfy only one side of this requirement. Voxel-projection methods establish explicit cross-view correspondence by lifting features into a shared 3D representation, but this introduces expensive voxel construction and often depends on object-specific 3D priors. Patch-attention methods, in contrast, avoid 3D construction and remain lightweight, but their fixed partitioning provides no explicit mechanism for cross-view geometry and may fragment defects that cross patch boundaries. We address this limitation with *Cross-view Deformable Fusion Module* (CDFM), a compact bottleneck that aligns multi-view features directly in 2D feature space, as shown in Figure 3. Given features from multiple views, CDFM learns data-dependent sampling offsets that specify where each query location should gather information from the remaining views. These offsets serve as an implicit geometric prior, enabling cross-view correspondence without camera calibration, voxel construction, or class-specific 3D supervision. Compared with fixed patch-wise interaction, CDFM introduces only a small overhead from offset prediction and bilinear sampling while providing geometry-aware deformable fusion.

For clarity, we describe CDFM at a single feature scale and omit the scale index k when no ambiguity arises. The same operation is applied independently at all selected feature scales. Let $\mathbf{T} \in \mathbb{R}^{B\times C\times H\times W}$ denote the teacher feature at the current scale, where $B = S\times V$. We reshape it into multi-view form $\mathbf{F} \in \mathbb{R}^{S\times V\times C\times H\times W}$. We use $q \in \{1,\dots,V\}$ to index the query view and $r \neq q$ to index a reference view.

Within-view spatial smoothing. Before cross-view fusion, we apply a parameter-free average pooling operation to reduce local discontinuities in the bottleneck features:

$$\mathbf{F}_{\text{sm}} = \text{AvgPool}_{7\times 7}(\mathbf{F}). \tag{2}$$

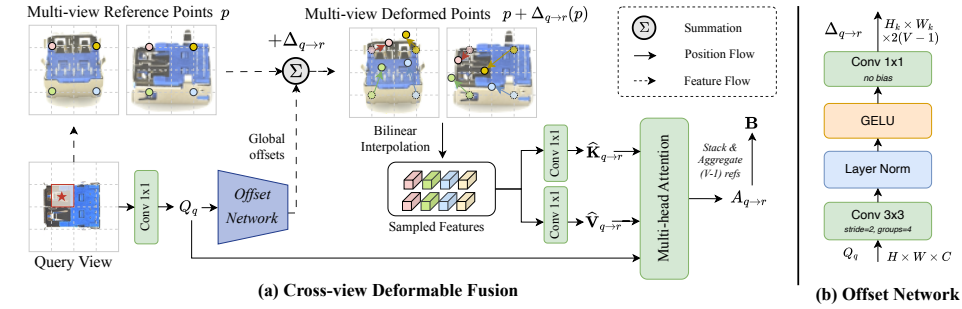

Figure 3: The mechanism of CDFM. (a) CDFM predicts offsets $\Delta_{q\to r}$ from the windowed query features and bilinearly samples reference keys and values at the deformed positions $\mathbf{p}+\Delta_{q\to r}$. Each query view produces $V-1$ offset fields, so windowed queries can deformably attend to features anywhere in each reference view, coupling cheap query windowing with global key/value reach. (b) Details of the offset network.

Here, $\text{AvgPool}_{7\times 7}$ denotes channel-wise average pooling with kernel size 7×7, stride 1, and padding 3, which preserves the spatial resolution. The operation is applied independently to each view and channel, introduces no learnable parameters, and smooths local boundary artifacts while preserving surface structure. We use the same kernel size at all feature scales and analyze this choice in Table 5.

Deformable cross-view attention. Given the smoothed multi-view feature $\mathbf{F}_{\text{sm}}$, CDFM treats each view as the query view in turn and uses the remaining views as references. The goal is to let each query location adaptively gather information from geometrically corresponding regions in other views, instead of attending to fixed patch windows. For a query view q and a reference view $r \neq q$, we compute the query, key, and value projections:

$$\mathbf{Q}_q = \text{Conv}_q(\mathbf{F}_{\text{sm},q}), \quad \mathbf{K}_r = \text{Conv}_k(\mathbf{F}_{\text{sm},r}), \quad \mathbf{V}_r = \text{Conv}_v(\mathbf{F}_{\text{sm},r}), \tag{3}$$

where Conv_q, Conv_k, and Conv_v are 1×1 convolutions for query, key, and value projection, respectively. To establish cross-view correspondence, CDFM predicts sampling offsets from the query feature rather than using fixed reference locations. A lightweight offset predictor generates one two-channel offset field for each reference view:

$$\Delta_q = \text{ConvOff}(\mathbf{Q}_q) \in \mathbb{R}^{S\times 2(V-1)\times H\times W}, \quad \Delta_{q\to r} = \Delta_q[r]. \tag{4}$$

Here, ConvOff is composed of a depthwise spatial convolution followed by a pointwise 1×1 convolution. Its output is divided into $V-1$ two-channel offset fields, where $\Delta_{q\to r}(p) \in \mathbb{R}^2$ denotes the learned displacement from query view q to reference view r at location p on the offset grid. Since different camera pairs induce different geometric displacements, predicting separate offsets for each reference view allows CDFM to model view-pair-specific correspondence while sharing the same offset predictor.

Using these offsets, reference keys and values are sampled at deformed positions by bilinear interpolation:

$$\hat{\mathbf{K}}_{q\to r}(p) = \mathbf{K}_r\left(p+\Delta_{q\to r}(p)\right), \quad \hat{\mathbf{V}}_{q\to r}(p) = \mathbf{V}_r\left(p+\Delta_{q\to r}(p)\right). \tag{5}$$

This deformable sampling step lets each query location retrieve reference-view features from continuous, content-adaptive positions, which alleviates the boundary fragmentation caused

by rigid patch partitioning. The sampled keys and values are then used to compute the cross-view attention response:

$$\mathbf{A}_{q\to r} = \mathrm{softmax}\left(\frac{\mathbf{Q}_q \left(\hat{\mathbf{K}}_{q\to r}\right)^\top}{\sqrt{d_h}}\right)\hat{\mathbf{V}}_{q\to r}, \tag{6}$$

where d_h is the per-head channel dimension. Finally, the fused bottleneck representation for each query view q is obtained by aggregating the responses from all reference views, and stacking across query views yields the full fused bottleneck at this scale:

$$\mathbf{B}_q = \frac{1}{V-1}\sum_{r\neq q}\mathbf{A}_{q\to r}, \quad \mathbf{B} = \mathrm{stack}_{q=1}^{V}\mathbf{B}_q \in \mathbb{R}^{S\times V\times C\times H\times W}. \tag{7}$$

We use uniform aggregation as a view-symmetric prior, since each camera provides a complementary observation of the same object surface. The learned offsets determine where each reference view contributes information, while the averaging step avoids introducing an additional view-selection module that may overfit to easy views or suppress difficult but informative ones.

Multi-scale geometric pyramid with global key-value reach. CDFM organizes the query side into a multi-scale pyramid: at each feature scale, we partition the query feature map into $n_{\text{win}} \times n_{\text{win}}$ non-overlapping windows and predict offsets within each window from the local query features. We set $n_{\text{win}} = \{8,4,2\}$ across the three feature scales, keeping the query window at a fixed 8×8 token size while letting its image-level receptive field grow from fine to coarse. Although query computation is window-partitioned for efficiency, the key/value grid covers the entire reference feature map. Sampling positions are expressed in image-normalized coordinates rather than window-local coordinates, so a query in any window can deformably attend to reference features anywhere in the reference view. This decoupled query-window / global-key-value design retains the cost benefit of windowed queries, $\mathcal{O}(|\text{window}|^2 \cdot n_{\text{windows}})$ instead of $\mathcal{O}((HW)^2)$ per view pair. At the same time, it avoids the artificial boundaries between windows that pure window partitioning would create, which would otherwise block correspondence across views.

3.3 Distributional View Alignment

CDFM learns local cross-view correspondence through deformable sampling, but local attention alone does not explicitly constrain the fused representations from different views to remain statistically consistent. We refer to this gap as *distributional inconsistency*: locally fused features may still exhibit misaligned global statistics across views. This matters for multi-view AD because different views of the same normal object can have different spatial layouts, scales, and appearances, yet should preserve compatible normality statistics. Enforcing pixel-level consistency is therefore inappropriate, since the same surface region may appear at different image locations across cameras. Instead, we regularize cross-view consistency in distribution space. We introduce *Distributional View Alignment* (DVA), a view-level regularization loss applied to the CDFM bottleneck features. DVA aligns the feature distributions of different views within the same object instance, encouraging view-consistent bottleneck representations without interfering with the teacher-student distillation objective and without requiring the memory or supervision of explicit 3D alignment.

Concretely, we align every view's bottleneck distribution against a per-instance *view-centric target* formed from the views themselves. This shared-target formulation avoids a

conflicting-gradient pathology of more direct pairwise alignment, where one badly-aligned pair can dominate the gradient and pull two views apart along an axis that other pairs are simultaneously aligning. A single centroid target removes this contention. Given bottleneck features $\mathbf{B}^k$ at scale k, let $\mathbf{b}_v^k$ denote the feature of view v. We first form the view-centric target $\bar{\mathbf{b}}^k$ by averaging across the V views within the same object instance:

$$\bar{\mathbf{b}}^k = \frac{1}{V}\sum_{v=1}^{V}\mathbf{b}_v^k. \tag{8}$$

We treat $\bar{\mathbf{b}}^k$ as a fixed target by stopping its gradient, denoted $\mathrm{sg}[\cdot]$ in Eq. 10. Without this, the averaging operation would propagate gradient from every view back into the centroid, creating a self-referential signal in which each view simultaneously shapes and is pulled toward its own target. We then convert each view feature and the centroid into probability distributions over the channel dimension via Softmax:

$$p_v^k = \mathrm{softmax}_C(\mathbf{b}_v^k), \quad \bar{p}^k = \mathrm{softmax}_C(\bar{\mathbf{b}}^k). \tag{9}$$

Distributional alignment is then enforced by the one-directional KL divergence of each view against the centroid, summed across views and scales:

$$\mathcal{L}_{DVA} = \sum_{k=1}^{K}\sum_{v=1}^{V} D_{KL}\left(p_v^k \,\middle\|\, \mathrm{sg}[\bar{p}^k]\right), \quad D_{KL}(p\,\|\,q) = \sum_c p(c)\log\frac{p(c)}{q(c)}. \tag{10}$$

where K and V denote the number of scales and views, respectively. The centric formulation is structurally symmetric across views (every view contributes equally to $\bar{p}^k$), so the asymmetric KL direction $D_{KL}(p_v^k\|\bar{p}^k)$ carries the design intent: it encourages each view to cover the modes of the consensus distribution rather than the consensus to cover any single view, which is the appropriate direction for normality modeling where capturing typical structure matters more than chasing outlier modes.

Assumption and scope. DVA constrains the cross-fusion bottleneck only during training; at inference, anomaly maps are derived from per-view teacher–student feature divergence, which DVA does not modify. Single-view anomaly evidence in the frozen teacher features therefore remains intact, ensuring that view-dependent anomalies are still recoverable through the distillation path.

Training objective. Our final training objective combines reconstruction fidelity with distributional consistency:

$$\mathcal{L}_{total} = \mathcal{L}_{KD} + \lambda\mathcal{L}_{DVA}, \tag{11}$$

where $\mathcal{L}_{KD}$ is the knowledge-distillation loss and λ is set to 1 to balance the two terms. $\mathcal{L}_{DVA}$ complements CDFM: CDFM establishes local cross-view correspondence through deformable sampling, while $\mathcal{L}_{DVA}$ enforces global distributional consistency on the resulting bottleneck. Together they yield a representation that is locally aligned and globally view-consistent without explicit 3D supervision.

4 Experiments

4.1 Experimental Settings

Datasets. Real-IAD [28] is a large-scale multi-view industrial AD dataset containing 30 object classes across diverse materials, including metal, plastic, and mixed materials, *etc.* The

Method	S-AUROC	S-AP	S-F1	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	P-AUPRO
RD4AD CVPR'22 [6]	78.3	63.7	68.3	80.5	59.5	62.4	91.8	29.1	34.7	78.8
UniAD NeurIPS'22 [51]	74.9	59.7	65.2	82.7	64.3	65.1	90.7	26.1	32.4	79.6
DeSTSeg CVPR'23 [57]	62.4	49.3	56.5	62.3	37.8	45.5	54.8	12.0	9.0	23.4
RealNet CVPR'24 [36]	55.1	40.9	51.7	57.3	31.8	41.3	50.0	10.3	4.4	15.1
MVAD TMM'25 [11]	<u>85.0</u>	<u>76.5</u>	<u>73.8</u>	<u>85.2</u>	<u>69.9</u>	<u>67.9</u>	<u>93.7</u>	<u>32.7</u>	<u>38.5</u>	<u>80.7</u>
GeoMAD (ours)	**88.2**	**81.9**	**78.3**	**87.0**	**73.4**	**70.5**	**94.2**	**36.8**	**41.6**	**83.1**

Table 2: Average performance comparison on MANTA-Tiny dataset. The best and second-best results are marked in **bold** and <u>underlined</u>.

Method	S-AUROC	S-AP	S-F1	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	P-AUPRO
RD4AD CVPR'22 [6]	85.7	92.4	87.7	83.0	79.6	74.4	97.3	25.8	33.5	90.4
UniAD NeurIPS'22 [51]	87.3	93.1	88.9	83.0	83.4	74.6	97.4	23.7	31.4	87.3
RD++ CVPR'23 [27]	87.0	93.3	88.2	83.7	80.9	74.9	97.7	25.8	33.5	90.0
SimpleNet CVPR'23 [17]	61.2	77.9	82.1	82.6	79.2	74.1	76.3	3.4	6.9	42.5
DeSTSeg CVPR'23 [57]	89.0	94.6	89.2	82.3	79.2	73.2	94.6	**37.9**	**41.7**	40.6
RealNet CVPR'24 [36]	72.0	87.1	82.5	65.2	65.4	62.7	60.5	27.2	24.9	28.7
MambaAD NIPS'24 [9]	85.7	78.3	74.7	85.7	72.2	69.1	93.8	35.5	40.6	80.7
ViTAD CVIU'25 [54]	88.8	94.3	89.3	82.8	80.0	73.7	97.1	23.9	31.8	84.3
MVAD TMM'25 [11]	<u>90.2</u>	<u>95.3</u>	<u>90.1</u>	<u>86.6</u>	<u>84.8</u>	<u>77.2</u>	**97.9**	30.3	36.8	<u>91.2</u>
GeoMAD (ours)	**91.8**	**95.9**	**91.2**	**87.5**	**85.6**	**78.2**	<u>97.8</u>	<u>35.6</u>	<u>40.8</u>	**92.2**

Table 3: Average performance comparison on Real-IAD dataset. The best and second-best results are marked in **bold** and <u>underlined</u>.

dataset comprises 150K high-resolution images with 99,721 normal and 51,329 abnormal samples. The training and testing sets follow standard splitting protocols, with normal samples used for training, and both normal and abnormal samples for testing. The dataset features eight distinct defect types (pit, deformation, scratch, *etc.*) with defect areas ranging from 0.1% to 6.75%, making it significantly more challenging than existing benchmarks. **MANTA** [7] is a visual-text AD dataset for tiny objects across 38 object classes spanning five typical domains (mechanics, medicine, electronics, agriculture, and food). The dataset comprises over 137.3K object instances, of which 8.6K objects are labeled as abnormal with pixel-level annotations. We evaluate on MANTA-Tiny, a subset of the MANTA dataset in which each class contains 800 object instances (4,000 individual images per class across 5 viewpoints), totaling 152,000 images across all 38 classes. Despite being termed "tiny," it contains more images than Real-IAD. All images in MANTA-Tiny are standardized to 256×256 resolution, facilitating consistent input preprocessing since GeoMAD processes only visual components. This contrasts with the original MANTA dataset, which requires manual cropping due to varying image dimensions and includes textual annotations that GeoMAD does not utilize.

Metrics. Following [11], we choose the metrics for evaluation across three levels of analysis. *Sample-level* metrics aggregate multi-view information to determine whether an object instance is abnormal, measured by S-AUROC, S-AP (Average Precision), and S-F1 (F1 score). *Image-level* metrics evaluate the detection capability for individual images using I-AUROC, I-AP, and I-F1. *Pixel-level* metrics assess the model's ability to localize abnormal regions precisely, including P-AUROC, P-AP, P-F1, and P-AUPRO (Area Under the Per-Region-Overlap). All results are the mean performance of all object classes for each dataset.

Implementation details. All input images from both the MANTA-Tiny and Real-IAD datasets are resized to 256×256 pixels without additional data augmentation. In the multi-view setting, we utilize all available view images with 8 input samples per batch, yielding a

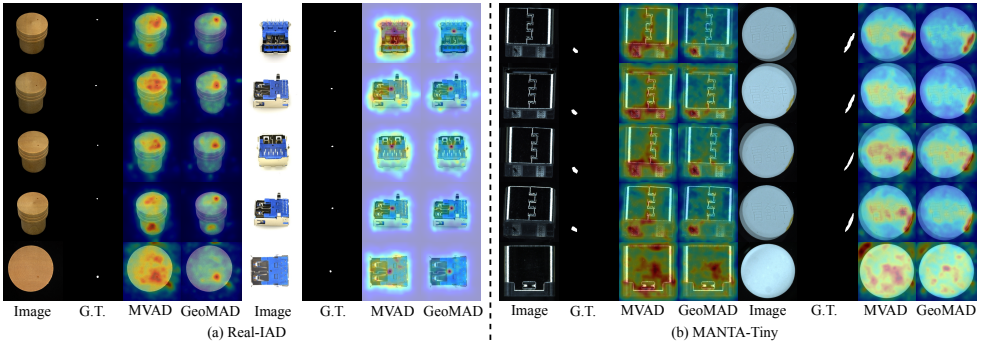

Figure 4: Qualitative comparison on Real-IAD (left) and MANTA-Tiny (right). Each row shows one view of the same object instance ($V = 5$ views per sample). GeoMAD yields more spatially coherent localization with fewer false activations.

total batch size of 40. Following RD4AD [6], we employ a teacher-student architecture in which the teacher network is a pre-trained ResNet-34 on ImageNet-1K that remains frozen during training, while the student network uses the same ResNet-34 architecture to learn to mimic the teacher's feature representations. Training is conducted for 50 epochs on a single NVIDIA RTX 5090 GPU. To ensure fair comparison, we reimplement prior methods on both the Real-IAD and MANTA-Tiny datasets in unified settings.

4.2 Comparison with the SOTA methods

We compare GeoMAD against a broad set of recent AD methods covering four design paradigms: reconstruction-based RD4AD [6] and RD++ [27], unified frameworks UniAD [31], MambaAD [9], and ViTAD [34], feature-based SimpleNet [17] and RealNet [36], segmentation-based DeSTSeg [37], and the multi-view method MVAD [11]. For fair comparison, all methods are configured in the multi-class setting, training a single model across all classes.

Results on Real-IAD. As shown in Table 3, GeoMAD achieves the best overall performance on Real-IAD, ranking first on 8 out of 10 metrics and second on the remaining two (P-AP and P-F1). Compared to the strongest reconstruction-based baseline MVAD [11], GeoMAD improves S-AUROC by 1.6, P-AP by 5.3, and P-F1 by 4.0 points, with consistent gains across sample-, image-, and pixel-level metrics. The only metrics on which GeoMAD does not lead are P-AP and P-F1, where DeSTSeg [37] ranks first. We note that DeSTSeg is trained with synthetic anomalies and pseudo ground-truth masks, providing direct pixel-level supervision for its segmentation head, whereas GeoMAD and the other baselines are trained only on anomaly-free data and derive anomaly maps from teacher–student divergence. This supervision asymmetry favors DeSTSeg on per-pixel metrics, but the advantage does not transfer to region-level localization (P-AUPRO) or to sample- and image-level detection.

Results on MANTA-Tiny. The advantage of GeoMAD grows on the more challenging MANTA-Tiny (Table 2), where high appearance variability makes multi-view reasoning harder. GeoMAD achieves the best result across all 10 metrics, surpassing the second-best MVAD by clear margins on every level: S-AUROC +3.2, S-AP +5.4, I-AP +3.5, P-AP +4.1, and P-AUPRO +2.4. We attribute these gains to CDFM's deformable cross-view fusion and DVA's distributional regularization, which together provide stronger inductive biases when each individual view carries limited or inconsistent evidence.

Qualitative comparison. Figure 4 shows qualitative comparisons against MVAD on both

Dataset	CDFM	DVA	S-AUROC	S-AP	S-F1	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	P-AUPRO
MANTA-Tiny	✗	✗	78.3	63.7	68.3	80.5	59.5	62.4	91.8	29.1	34.7	78.9
	✓	✗	87.8	81.4	77.7	86.6	72.7	70.0	94.0	36.0	40.9	82.4
	✓	✓	**88.2**	**81.9**	**78.3**	**87.0**	**73.4**	**70.5**	**94.2**	**36.8**	**41.6**	**83.1**
Real-IAD	✗	✗	84.7	91.8	87.5	82.0	78.5	73.7	97.0	26.1	33.3	89.5
	✓	✗	91.6	95.8	91.0	87.3	85.5	77.9	**97.8**	**35.7**	**41.0**	92.0
	✓	✓	**91.8**	**95.9**	**91.2**	**87.5**	**85.6**	**78.2**	**97.8**	35.6	40.8	**92.2**

Table 4: Ablation study on CDFM and DVA. We evaluate the contribution of Cross-view Deformable Fusion Module (CDFM) and Distributional View Alignment (DVA) on MANTA-Tiny and Real-IAD, reporting sample-level (S), image-level (I), and pixel-level (P) metrics. **Bold** indicates the best result and underlined indicates the second best.

datasets. MVAD's patch-based fusion tends to fragment defects near patch boundaries and over-activate on background regions, while `GeoMAD` produces spatially coherent localization that remains consistent across the five views of the same object instance. The behavior reflects CDFM's deformable cross-view sampling and DVA's distributional regularization, and aligns with our quantitative results: `GeoMAD` achieves notably higher P-AP and P-F1 than MVAD, confirming that resolving patch-boundary fragmentation directly reduces false positive predictions. The gap is more pronounced on MANTA-Tiny, where MVAD produces diffuse activations across the object surface while `GeoMAD` retains focused localization.

4.3 Ablations

Component Analysis. We assess the individual contributions of CDFM and DVA in Table 4 by incrementally adding each component to a baseline that uses neither. Adding CDFM yields substantial gains on both datasets, with particularly large improvements on the more challenging MANTA-Tiny benchmark. This confirms that deformable cross-view fusion provides a strong inductive bias for multi-view AD, allowing the model to learn view correspondence directly from data without explicit 3D supervision. Adding DVA on top of CDFM further improves performance across the majority of metrics on both datasets, with the gains being most consistent on MANTA-Tiny. While the improvement from DVA is smaller than that of CDFM, this is expected: CDFM establishes local cross-view correspondence, leaving DVA the complementary role of regularizing the resulting bottleneck distributions to be globally view-consistent. Together, the two components yield the best overall performance on both datasets, validating their complementary design.

Smoothing Kernel Size Analysis. We study how the receptive field of the smoothing kernel in CDFM affects anomaly detection performance, comparing single-kernel sizes from 3 to 9 and a multi-kernel variant $\{5,7\}$ against a no-smoothing baseline ("-"). As shown in Table 5, smoothing is beneficial when the receptive field is large enough: kernel sizes 5, 7, and 9 all outperform the no-smoothing baseline, while the smallest kernel of size 3 actually underperforms it, suggesting that an overly local average disrupts useful structure without effectively denoising boundary artifacts. Among the effective kernels, size 7 provides the best overall balance, achieving the strongest results on the majority of image- and pixel-level metrics. Larger kernels such as 9 remain competitive on sample-level metrics but slightly degrade pixel-level localization, indicating that excessive smoothing blurs fine-grained anomaly cues. Interestingly, combining multiple kernels $\{5,7\}$ does not yield gains and instead performs on par with kernel 3, suggesting that a single well-chosen receptive field is sufficient for the smoothing operation. We therefore adopt kernel size 7 as the default in all

Kernel Size	S-AUROC	S-AP	S-F1	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	P-AUPRO
-	91.3	95.7	90.9	87.0	85.1	77.6	**97.8**	35.5	40.5	91.6
{3}	91.0	95.5	90.7	86.8	85.0	77.4	**97.8**	35.2	40.3	91.4
{5}	91.6	95.8	91.0	87.4	85.5	78.0	**97.8**	**36.1**	**41.2**	92.0
{7}	91.8	95.9	**91.2**	**87.5**	**85.6**	**78.2**	**97.8**	35.6	40.8	**92.2**
{9}	**92.0**	**96.0**	**91.2**	87.3	85.3	77.9	**97.8**	35.1	40.5	92.0
{5,7}	91.0	95.5	90.7	86.8	84.9	77.4	**97.8**	35.4	40.5	91.4

Table 5: Impact of smoothing kernel size in CDFM. We compare single-kernel sizes {3, 5, 7, 9} and a multi-kernel variant {5, 7} against a no-smoothing baseline (denoted "-"). The best and second-best results are marked in **bold** and underlined.

DVA Type	S-AUROC	S-AP	S-F1	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	P-AUPRO
View Pair	91.6	95.8	91.2	86.9	84.8	77.4	97.8	34.3	39.8	91.7
View Centric	**91.8**	**95.9**	**91.2**	**87.5**	**85.6**	**78.2**	**97.8**	**35.6**	**40.8**	**92.2**

Table 6: Alignment target in DVA. Comparison between pairwise alignment (View Pair) and view-centric alignment toward a shared per-instance centroid (View Centric).

subsequent experiments.

Impact of Missing Views. A practical concern in multi-view deployment is robustness when certain camera views become unavailable due to occlusion or hardware failure. To assess this, we take the model trained with the default 5 views and evaluate it with progressively fewer views (5→2) at inference time. As shown in Table 7, results are reported at the Sample (S), Image (I), and Pixel (P) levels. Image- and pixel-level performance remain highly stable as views are removed: I-AUROC drops by only 0.8 points and P-AUROC by 0.7 points even when 60% of views are unavailable, indicating that per-view detection and localization do not depend on the presence of any specific view. Sample-level metrics show a larger but graceful decline, which is structurally expected since the sample score aggregates evidence across views and naturally weakens as the number of views decreases. Overall, the method retains strong per-view performance under substantial view loss, supporting its deployment in scenarios where capturing all views is not always feasible.

Alignment Target in DVA. A central design choice in DVA is what each view should be aligned against. We compare two natural targets: *View Pair*, which enforces consistency between every pair of view distributions via symmetric KL divergence ($\mathcal{O}(V^2)$ constraints), and *View Centric*, which aligns each view against a shared per-instance centroid formed from the views themselves ($\mathcal{O}(V)$ constraints, Eq. 8). As shown in Table 6, View Centric consistently outperforms View Pair across all metrics, with the most pronounced gains on fine-grained pixel-level metrics. Beyond the lower constraint count, the view-centric formulation avoids a conflicting-gradient pathology of pairwise objectives, where a single badly-aligned pair can dominate the gradient and pull two views apart along an axis that other pairs are simultaneously aligning. A shared centroid target removes this contention by giving all views one symmetric reference, providing a more coherent learning signal.

Computational Efficiency. Table 8 compares the computational cost of `GeoMAD` with representative SOTA methods, measured with batch size 40. `GeoMAD` maintains a compact footprint of 19.0M parameters, on par with the most parameter-efficient baseline MVAD [11], while using 4.2× fewer parameters and 3.0× lower FLOPs than RD4AD [6]. Although UniAD [31] achieves the lowest FLOPs, `GeoMAD` offers a more favorable balance across all three dimensions, particularly compared to MambaAD [9], whose training memory is over 3× higher than ours. Notably, IDIF is estimated to require over 149M parameters based on

Views	S-AUROC	S-AP	S-F1	I-AUROC	I-AP	I-F1	P-AUROC	P-AP	P-F1	P-AUPRO
5 (default)	**91.8**	**95.9**	**91.2**	**87.5**	**85.6**	**78.2**	**97.8**	**35.6**	**40.8**	**92.2**
4	90.9	95.5	90.4	87.2	85.3	78.0	97.7	35.5	**40.8**	91.9
3	89.1	94.7	89.1	86.9	84.8	77.5	97.5	35.3	40.6	91.8
2	85.5	92.9	87.3	86.7	83.4	76.5	97.1	34.2	39.6	91.5

Table 7: Robustness to missing views on the Real-IAD dataset. We evaluate model performance when reducing the number of available views from the default 5 down to 2 at inference time, using the model trained with all 5 views.

Method	Parameters	FLOPs	Train Memory
RD4AD $_{\text{CVPR'22}}$ [8]	80.61M	142.0G	11,614M
UniAD $_{\text{NeurIPS'22}}$ [31]	24.52M	**16.8G**	7,324M
ViTAD $_{\text{CVIU'25}}$ [34]	38.6M	48.3G	**4,842M**
MambaAD $_{\text{NIPS'24}}$ [9]	25.7M	41.5G	28,994M
MVAD $_{\text{TMM'25}}$ [11]	**18.4M**	45.0G	6,744M
GeoMAD (ours)	19.0M	46.8G	8,310M

Table 8: Computational efficiency comparison. Parameters, FLOPs, and training memory of GeoMAD against representative SOTA methods.

its voxel-related modules and comparable architectures [30, 37], underscoring the efficiency of our non-voxel design.

5 Conclusion

We propose GeoMAD, a unified model for multi-view, multi-class anomaly detection that addresses two complementary limitations: geometric correspondence deficiency and distributional inconsistency. For cross-view geometric correspondence, we propose *Cross-view Deformable Fusion Module* (CDFM), a calibration-free 2D bottleneck design that learns content-adaptive, view-pair-specific sampling offsets with image-global reach across a multi-scale window pyramid. For distributional consistency, we introduce *Distributional View Alignment* (DVA), which aligns each view's bottleneck distribution against a stop-gradient view-centric target, requiring no anomaly labels, pixel-level matching, or 3D supervision. Together, CDFM and DVA combine local deformable correspondence with global distributional consistency in a single 2D architecture. Experiments on Real-IAD and MANTA-Tiny show that GeoMAD improves unified multi-view anomaly detection and localization over existing methods, supporting the value of jointly addressing geometric and distributional gaps in cross-view learning. Future work could explore extending GeoMAD to handle dynamic view configurations.

Acknowledgment: The work of Shang-Fu Chen and Wen-Huang Cheng was supported in part by the National Science and Technology Council, Taiwan (Grants: NSTC-112-2628-E-002-033-MY4 and NSTC-114-2634-F-002-004), the Taiwan Centers of Excellence (TCE), and the Center of Data Intelligence: Technologies, Applications, and Systems (Grants: 115L900901 / 115L900902 / 115L900903), National Taiwan University, from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education, Taiwan. The work of Kai-Lung Hua was supported in part by the National Science and Technology Council, Taiwan (Grants: NSTC-114-2221-E-011-045-MY3 and NSTC-114-2221-E-011-041-MY3). Kuan-Chuan Peng was exclusively supported by MERL.

References

[1] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. MVTec AD–a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[2] Yu Cai, Hao Chen, and Kwang-Ting Cheng. Rethinking autoencoders for medical anomaly detection from a theoretical perspective. In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024.

[3] Yu Cai, Weiwen Zhang, Hao Chen, and Kwang-Ting Cheng. MedIAnomaly: A comparative study of anomaly detection in medical images. *Medical Image Analysis*, page 103500, 2025.

[4] Zining Chen, Xingshuang Luo, Weiqiu Wang, Zhicheng Zhao, Fei Su, and Aidong Men. Filter or compensate: Towards invariant representation from distribution shift for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[5] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. PaDiM: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition (ICPR)*, 2021.

[6] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[7] Lei Fan, Dongdong Fan, Zhiguang Hu, Yiwen Ding, Donglin Di, Kai Yi, Maurice Pagnucco, and Yang Song. MANTA: A large-scale multi-view and visual-text anomaly detection dataset for tiny objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[8] Jia Guo, Shuai Lu, Weihang Zhang, Fang Chen, Huiqi Li, and Hongen Liao. Dinomaly: The less is more philosophy in multi-class unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[9] Haoyang He, Yuhu Bai, Jiangning Zhang, Qingdong He, Hongxu Chen, Zhenye Gan, Chengjie Wang, Xiangtai Li, Guanzhong Tian, and Lei Xie. MambaAD: Exploring state space models for multi-class unsupervised anomaly detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[10] Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[11] Haoyang He, Jiangning Zhang, Guanzhong Tian, Chengjie Wang, and Lei Xie. Learning multi-view anomaly detection with efficient adaptive selection. *IEEE Transactions on Multimedia (TMM)*, 2025.

[12] Teng Hu, Jiangning Zhang, Ran Yi, Yuzhen Du, Xu Chen, Liang Liu, Yabiao Wang, and Chengjie Wang. AnomalyDiffusion: Few-shot anomaly image generation with diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

[13] Fei Li, Wenxuan Liu, Jingjing Chen, Ruixu Zhang, Yuran Wang, Xian Zhong, and Zheng Wang. Anomize: Better open vocabulary video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[14] Yuxuan Lin, Yang Chang, Xuan Tong, Jiawen Yu, Antonio Liotta, Guofan Huang, Wei Song, Deyu Zeng, Zongze Wu, Yan Wang, et al. A survey on rgb, 3d, and multimodal approaches for unsupervised industrial image anomaly detection. *Information Fusion*, 2025.

[15] Jing Liu, Yang Liu, Jieyu Lin, Jielin Li, Liang Cao, Peng Sun, Bo Hu, Liang Song, Azzedine Boukerche, and Victor CM Leung. Networking systems for video anomaly detection: A tutorial and survey. *ACM Computing Surveys (CSUR)*, 57(10):1–37, 2025.

[16] Xinyue Liu, Jianyuan Wang, Biao Leng, and Shuo Zhang. Unlocking the potential of reverse distillation for anomaly detection. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[17] Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. SimpleNet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[18] Wenxi Lv, Qinliang Su, and Wenchao Xu. One-for-all few-shot anomaly detection via instance-induced prompt learning. In *International Conference on Learning Representations (ICLR)*, 2025.

[19] Snehashis Majhi, Giacomo D'Amicantonio, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Egor Bondarev, and François Brémond. Just dance with pi! a poly-modal inductor for weakly-supervised video anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[20] Kai Mao, Yiyang Lian, Yangyang Wang, Meiqin Liu, Nanning Zheng, and Ping Wei. Unveiling multi-view anomaly detection: Intra-view decoupling and inter-view fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[21] Atsuyuki Miyai, Jingkang Yang, Jingyang Zhang, Yifei Ming, Yueqian Lin, Qing Yu, Go Irie, Shafiq R. Joty, Yixuan Li, Hai Li, Ziwei Liu, T. Yamasaki, and Kiyoharu Aizawa. Generalized out-of-distribution detection and beyond in vision language model era: A survey. *Transactions on Machine Learning Research (TMLR)*, 2025.

[22] Mojtaba Nafez, Amirhossein Koochakian, Arad Maleki, Jafar Habibi, and Mohammad Hossein Rohban. PatchGuard: Adversarially robust anomaly detection and localization through vision transformers and pseudo anomalies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[23] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Computing Surveys (CSUR)*, 54(2): 1–38, 2021.

[24] Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[25] Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International Conference on Machine Learning (ICML)*, 2018.

[26] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection. In *International Conference on Learning Representations (ICLR)*, 2020.

[27] Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan Duong, Chanh D Tr Nguyen, and Steven QH Truong. Revisiting reverse distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[28] Chengjie Wang, Wenbing Zhu, Bin-Bin Gao, Zhenye Gan, Jiangning Zhang, Zhihao Gu, Shuguang Qian, Mingang Chen, and Lizhuang Ma. Real-IAD: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[29] Shun Wei, Jielin Jiang, and Xiaolong Xu. UniNet: A contrastive learning-guided unified framework with feature selection for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[30] Haozhe Xie, Hongxun Yao, Xiaoshuai Sun, Shangchen Zhou, and Shengping Zhang. Pix2vox: Context-aware 3d reconstruction from single and multi-view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[31] Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[32] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. DRAEM-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[33] Chongsheng Zhang, George Almpanidis, Gaojuan Fan, Binquan Deng, Yanbo Zhang, Ji Liu, Aouaidjia Kamel, Paolo Soda, and João Gama. A systematic review on long-tailed learning. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2025.

[34] Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. Exploring plain vit features for multi-class unsupervised visual anomaly detection. *Computer Vision and Image Understanding (CVIU)*, 253:104308, 2025.

[35] Ximiao Zhang, Min Xu, Dehui Qiu, Ruixin Yan, Ning Lang, and Xiuzhuang Zhou. Mediclip: Adapting clip for few-shot medical image anomaly detection. In *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2024.

[36] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. RealNet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[37] Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, and Ting Chen. DeSTSeg: Segmentation guided denoising student-teacher for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[38] Liyun Zhu, Lei Wang, Arjun Raj, Tom Gedeon, and Chen Chen. Advancing video anomaly detection: A concise review and a new dataset. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[39] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision (ECCV)*, 2022.