# RAID: Retrieval-Augmented Anomaly Detection

Mingxiu Cai[1,*] Zhe Zhang[1,*] Gaochang Wu[1,†] Tianyou Chai[1] Xiatian Zhu[2]
[1]State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University
[2]University of Surrey

## Abstract

*Unsupervised Anomaly Detection (UAD) aims to identify abnormal regions by establishing correspondences between test images and normal templates. Existing methods primarily rely on image reconstruction or template retrieval but face a fundamental challenge: matching between test images and normal templates inevitably introduces noise due to intra-class variations, imperfect correspondences, and limited templates. Observing that Retrieval-Augmented Generation (RAG) leverages retrieved samples directly in the generation process, we reinterpret UAD through this lens and introduce **RAID**, a retrieval-augmented UAD framework designed for noise-resilient anomaly detection and localization. Unlike standard RAG that enriches context or knowledge, we focus on using retrieved normal samples to guide noise suppression in anomaly map generation. RAID retrieves class-, semantic-, and instance-level representations from a hierarchical vector database, forming a coarse-to-fine pipeline. A matching cost volume correlates the input with retrieved exemplars, followed by a guided Mixture-of-Experts (MoE) network that leverages the retrieved samples to adaptively suppress matching noise and produce fine-grained anomaly maps. RAID achieves state-of-the-art performance across full-shot, few-shot, and multi-dataset settings on MVTec, VisA, MPDD, and BTAD benchmarks.* https://github.com/Mingxiu-Cai/RAID.

## 1. Introduction

Anomaly detection serves as a cornerstone task in computer vision with applications in industrial quality inspection [38, 56, 66], medical image analysis [25], and intelligent surveillance [58]. Given the limited availability and diverse nature of anomaly patterns, recent studies [2, 15, 21] have increasingly gravitated toward Unsupervised Anomaly Detection (UAD) without access to anomalous samples. In parallel, the shift from the traditional one-class-one-model paradigm [49, 63] toward unified multi-class UAD [59, 60, 71] is steadily enhancing its practical applicability.

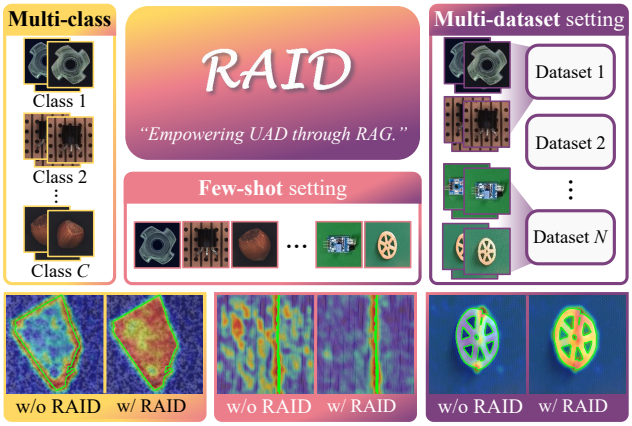


Figure 1. We reformulate UAD within the Retrieval-Augmented Generation (RAG) paradigm, effectively reducing retrieval and matching noise and enabling stronger generalization across full-shot, few-shot, and multi-dataset settings.

Existing UAD methods can be broadly categorized into reconstruction- and embedding-based paradigms [36]. Reconstruction-based methods leverage Generative Adversarial Networks (GANs) [24], Transformers [21, 41, 64], or diffusion models [63] to learn the manifold of normal patterns, projecting inputs onto this manifold to obtain normal-looking reconstructions, where discrepancies from the inputs reveal anomalies. Embedding-based methods [45, 55, 67], in contrast, bypass explicit reconstruction and instead perform feature-level matching between a query and its corresponding normal templates stored in a memory bank, which may consist of image-level features [18], or patch-level embeddings [11, 47, 49]. Recently, growing attention has been directed toward leveraging vision foundation models [46, 48, 52] to enhance both paradigms with semantically rich representations, e.g., WinCLIP [26], AnomalyDINO [11], and Dinomaly [21]. However, these approaches still suffer from unreliable feature matching arising from imperfect reconstructions [36] or suboptimal template retrieval, and limited few-shot generalization when adapting to domain-specific categories [1].

In this paper, we reinterpret UAD through the lens of the Retrieval-Augmented Generation (RAG) paradigm, as illus-

*Equal Contribution.
†Gaochang Wu (wugc@mail.neu.edu.cn) is the corresponding author.

trated in Fig. 1. In the era of reasoning-driven artificial intelligence, RAG has emerged as an effective framework to mitigate hallucinations and enhance generalization when models lack sufficient domain-specific knowledge or data [17]. It has been successfully applied to diverse vision tasks [72], including conditional image generation [5], long-tailed image classification [40], and open-vocabulary object detection [28]. From this perspective, most reconstruction- and embedding-based UAD approaches can be conceptualized as partial realizations of the RAG pipeline, where the model `retrieves` a normal counterpart (via reconstruction, e.g., GLAD [59], memory retrieval, e.g., PatchCore [49], or teacher-student distillation, e.g., RD++ [54]) and identifies anomalies via feature matching. Despite this conceptual alignment, most existing methods overlook the `generative` reasoning stage of RAG, producing hallucinatory detection noise (e.g., blurred anomaly boundaries and missing subtle anomalies) due to unreliable feature matching.

To address this gap, we propose RAID, a Retrieval-Augmented Industrial anomaly Detection framework that fully integrates the RAG pipeline for UAD, which `retrieves` normal representations and subsequently `generates` anomaly maps by jointly reasoning over the retrieved patches and the input patches. To achieve efficient and scalable retrieval, we design a hierarchical vector database that organizes tokenized templates into three levels of entities, class prototype (category-level concept), semantic prototype (clustered patch token), and instance token (patch token), rather than adopting a flat structure. Compared to existing image-level template retrieval and flat retrieval schemes, this hierarchy facilitates a coarse-to-fine retrieval flow, enabling query tokens to efficiently access semantically relevant template patches with strong contextual consistency for downstream anomaly generation.

We model the generation stage as a guided Mixture-of-Experts (MoE) filtering network designed to mitigate potential matching noise between the input and its multiple retrieved counterparts. Following CostFilter-AD [68], we first construct a matching cost volume by correlating each input token with its retrieved template tokens. In contrast to previous approaches [11, 68], the hierarchical retrieval structure in RAID effectively reduces the matching dimensionality in the cost volume while preserving semantic fidelity. To further mitigate unreliable feature correspondence, the proposed guided MoE filter leverages both the input tokens and the retrieved semantic prototypes as dual guidance, adaptively assigning multiple denoising experts that specialize in distinct semantic and spatial contexts to generate a refined anomaly map that preserves fine-grained anomaly boundaries and subtle anomalies. Leveraging category- and dataset-agnostic retrieval, RAID injects universal semantic priors into a dynamically activated MoE filter, enabling it to focus on matching-cost denoising and learn category-agnostic anomaly representations. This joint retrieval-generation scheme leads to robust anomaly localization and strong few-shot generalization across unseen categories, as demonstrated in Fig. 1 (bottom).

We summarize our contributions as follows: 1) We propose RAID, a new paradigm that reconceptualizes UAD within the RAG framework, enabling reliable detection and localization with strong generalization. 2) We introduce a hierarchical vector database that enables a coarse-to-fine retrieval flow, allowing query tokens to efficiently access semantically relevant template tokens with strong contextual consistency. 3) We design the generation stage as a guided MoE filtering network, which dynamically allocates denoising experts to suppress hallucinatory matching noise and improve robustness against diverse anomaly distributions. 4) We extensively evaluate RAID under full-shot, few-shot, and multi-dataset settings, where it consistently outperforms existing methods across multiple benchmarks, demonstrating superior generalization and scalability.

## 2. Related works

### 2.1. Unsupervised Anomaly Detection

Early studies mainly followed the *single-class* paradigm, where models are trained on normal samples to detect anomalies at test time [49, 62]. With the advent of foundation models and the emergence of new architectures [12, 29, 46, 52], research has progressed along two directions: (i) *multi-class* detection using a unified model to identify all categories within a dataset [34, 41, 59, 60, 68]; and (ii) *few-shot detection* for unknown classes unseen during training [7, 9, 69, 73], which leverages a small set of auxiliary anomalies without class overlap to the test domain. Existing approaches can be broadly grouped into *reconstruction-based* and *embedding-based* families [6, 36]. Reconstruction-based pipelines attempt to map anomalous regions back to normal ones and identify discrepancies through similarity or residual analysis. Representative architectures include U-Net [62, 71], Vision Transformer (ViT) [20, 21], Mamba [22], Diffusion [23, 34], and MoE-based [43] frameworks. These models typically integrate pre-trained features with customized decoders [21, 22, 41, 59, 60] to address the pervasive "identical shortcut" issue [36], which weakens matching similarity or residual contrast.

Embedding-based methods learn either class-specific [39, 49, 67] or class-agnostic [47, 55] normal representations, and detect outliers by modeling a normal distribution [74], maintaining a memory bank [49] that stores image-level [18] or patch-level embeddings [11, 47, 49]. While effective, these methods often suffer from high computational overhead due to large

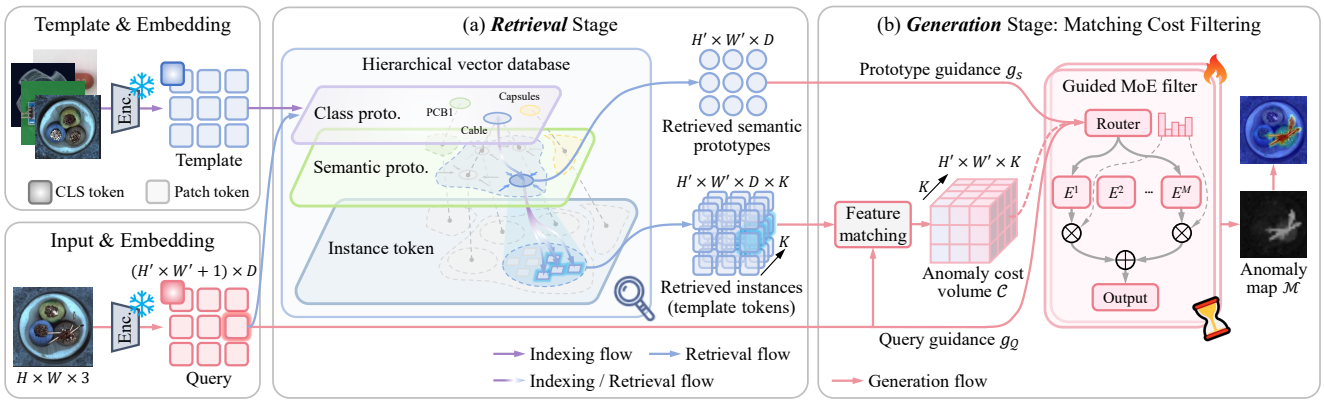


Figure 2. Overview of our RAID, which reinterprets UAD within a RAG paradigm. (a) In the `retrieval` stage, a hierarchical vector database is constructed, indexing tokenized templates into three sequential entity levels: class prototype, semantic prototype, and instance token. This structure allows efficient retrieval flow queried by input tokens. (b) In the `generation` stage, an anomaly cost volume is built by matching each input token with its retrieved template tokens. A guided MoE filter then dynamically refines this cost volume under the dual guidance of the retrieved semantic prototypes and the input tokens.

template search spaces [36] or matching noise caused by non-ideal retrieval templates. CostFilter-AD [68] further introduces a matching-cost filtering module as a plug-in to suppress such noise, yet its performance remains constrained by the initial anomaly cues provided by the host models. Recently, CLIP-based approaches [10, 26, 42, 73] have been explored for novel-class anomaly detection in the *few-shot* setting, where text embeddings are treated as semantic templates. However, their success heavily depends on the diversity and representativeness of auxiliary real anomaly data.

### 2.2. Retrieval-Augmented Generation in Vision

Retrieval-Augmented Generation (RAG) injects retrievable external knowledge around encoding/decoding to align with model context [17, 32, 50], improving factual consistency, interpretability, and timeliness. Originally developed for language and text understanding [17], RAG has progressively extended to vision tasks [72]. In image understanding, notable directions include retrieval-augmented open-vocabulary object detection [28], image captioning and generation [51], medical image segmentation for clinical decision-making [70], and question answering over visually rich, mixed-format documents [53]. For visual generation, retrieved image priors are integrated into diffusion and reconstruction pipelines to refine structural and textural fidelity, exemplified by retrieval-augmented text-to-image generation [8], image restoration [19], and super-resolution [31]. In video understanding, CadenceRAG [37] demonstrates that retrieval across large video libraries enhances text-grounded video retrieval and question answering, while multi-query and multi-evidence strategies further advance multimodal RAG [30, 61].

A conceptually related work is ADSeeker [65], which casts anomaly analysis as vision and language question answering that retrieves image-document knowledge to condition an MLLM [3] for generating textual anomaly descriptions and image-level decisions. In contrast, RAID targets purely visual UAD with pixel-level output (i.e., segmentation) without relying on text Information. It performs hierarchical retrieval over visual exemplars to strike a balance between accuracy and efficiency, and further leverages filtering-based generative reasoning to suppress matching noise and enhance anomaly localization. This design not only clearly distinguishes RAID from existing retrieval-only approaches [11, 18, 47, 49], but also delivers robust generalization and superior performance across *multi-class*, *few-shot*, and *multi-dataset* UAD settings.

## 3. Methodology

### 3.1. Overview

Given an input (query) image $x_{\mathcal{Q}} \in \mathbb{R}^{H\times W\times 3}$ (channel, height, and width) or a template (reference) image $x_{\mathcal{T}} \in \mathbb{R}^{H\times W\times 3}$ prepared for the database, a pre-trained ViT encoder is employed to embed them into patch tokens $\{\mathbf{p}_{\mathcal{Q}}\}, \{\mathbf{p}_{\mathcal{T}}\} \in \mathbb{R}^{(H'\times W')\times D}$ and class (CLS) tokens $\mathbf{c}_{\mathcal{Q}}$, $\mathbf{c}_{\mathcal{T}} \in \mathbb{R}^{1\times D}$, where $H'$, $W'$, and $D$ denote the resulting spatial resolution and token dimension, respectively. The cosine similarity between a query token vector $\mathbf{p}_{\mathcal{Q}}$ and a template token $\mathbf{p}_{\mathcal{T}}$ is defined as:

$$\mathrm{sim}(\mathbf{p}_{\mathcal{Q}}, \mathbf{p}_{\mathcal{T}}) = \frac{\mathbf{p}_{\mathcal{Q}}\mathbf{p}_{\mathcal{T}}^{\top}}{\|\mathbf{p}_{\mathcal{Q}}\|\,\|\mathbf{p}_{\mathcal{T}}\|}, \tag{1}$$

where $\|\cdot\|$ is the $L_2$ norm. In most existing retrieval-only approaches, both template retrieval and anomaly scoring rely on this similarity measure. However, directly searching across all template tokens in the database incurs substantial computational overhead and generalizes poorly to unseen categories. Moreover, when the number of templates is lim-

ited, as in few-shot scenarios, the detection performance becomes highly sensitive to noisy or unreliable templates.

Motivated by the RAG paradigm [17, 72], which alleviates hallucinations through augmented generation, we enhance UAD by further introducing a filtering-based generation reasoning process, as illustrated in Fig. 2. The overall RAID framework can be formulated as:

$$\mathcal{M} = \mathcal{G}(\{\mathbf{p}_{\mathcal{Q}}\}, \mathcal{R}(\{\mathbf{p}_{\mathcal{Q}}\}, \mathcal{D})). \tag{2}$$

Here, $\mathcal{R}(\{\mathbf{p}_{\mathcal{Q}}\}, \mathcal{D})$ performs hierarchical retrieval from the database $\mathcal{D}$, balancing search efficiency and retrieval accuracy; $\mathcal{G}(\cdot,\cdot)$ then executes guided MoE filtering to generate the refined anomaly map $\mathcal{M}$, conditioned on the query and retrieved templates, effectively suppressing retrieval noise.

### 3.2. Hierarchical Vector Database Construction

Existing UAD methods typically adopt a flat retrieval structure [11, 18, 49], where each query patch searches for its globally most similar counterpart in a large memory bank. Such a design leads to high computational overhead and degraded inference efficiency as the bank size grows. To balance retrieval accuracy and efficiency while ensuring inter-class discriminability and intra-class representational richness, we introduce a hierarchical vector database $\mathcal{D}$, as illustrated in Fig. 2(a), with the following `indexing flow` of template tokens: class prototype $\{\bar{\mathbf{c}}\} \rightarrow$ semantic prototype $\{\bar{\mathbf{s}}\} \rightarrow$ instance token $\{\bar{\mathbf{t}}\}$. During database construction, both the CLS token $\mathbf{c}_{\mathcal{T}}$ and the patch tokens $\{\mathbf{p}_{\mathcal{T}}\}$ are utilized to form this multi-level hierarchy.

The **class prototype entity** indexes template tokens into corresponding class-level clusters. To obtain it, we perform K-means clustering on the CLS tokens from all templates:

$$\{\bar{\mathbf{c}}\} = \mathrm{KMeans}\big(\{\mathbf{c}_{\mathcal{T}}^{n}\}_{n=1}^{N}\big),$$

where $\mathrm{KMeans}(\cdot)$ denotes the K-means operation, $N$ is the number of templates, and $\bar{\mathbf{c}} \in \mathbb{R}^{1\times D}$ represents the class prototype (class-level centroids), with $|\{\bar{\mathbf{c}}\}| = C$ as the class number. Leveraging the class-level semantics encoded in CLS tokens, the class prototype entity enables category- and dataset-agnostic retrieval, providing a strong scalability capability for multi-dataset UAD.

The **semantic prototype entity** organizes template tokens into an intra-class semantic structure:

$$\{\bar{\mathbf{s}}\}^{c} = \mathrm{KMeans}\big(\{\mathbf{p}_{\mathcal{T}}^{c,n}\}_{n=1}^{N^{c}}\big),$$

where $\bar{\mathbf{s}}^{c} \in \mathbb{R}^{1\times D}$ represents the resulting semantic prototype. Note that multiple semantic prototypes are generated per class, and we use $\bar{\mathbf{s}}^{c,j}$ to indicate the $j$-th semantic prototype guided by the class prototype entity $\bar{\mathbf{c}}^{c}$. These prototypes capture recurring intra-class patterns, such as textures, structural components, or backgrounds, and serve as structured guidance for more effective anomaly detection.

The **instance token entity** stores all template tokens $\{\mathbf{p}_{\mathcal{T}}\}$ in organized instance token sets $\{\mathbf{t}\}^{c,j}$, sequentially indexed following the flow defined by the class prototype and semantic prototype entities. Here, we use $\mathbf{t}^{c,j,k} \in \mathbb{R}^{1\times D}$ to denote the $k$-th instance token associated with the semantic prototype $\bar{\mathbf{s}}^{c,j}$. This entity preserves fine-grained visual details, enabling accurate retrieval and pixel-level matching for downstream anomaly detection.

### 3.3. Hierarchical Retrieval

The hierarchical vector database naturally supports a coarse-to-fine `retrieval flow`, as demonstrated in Fig. 2(a), which refines correspondences by narrowing the search space from global class-level clusters to intra-class semantics, and finally to local instance-level template tokens. This hierarchical retrieval mechanism ($\mathcal{R}$ in Eqn. (2)) effectively reduces redundant matching and mitigates scalability bottlenecks observed in prior works [11, 68], thereby enhancing the applicability of the proposed RAID framework to large-scale datasets.

At the top level, the CLS token of the input $\mathbf{c}_{\mathcal{Q}}$ is compared with the class prototypes $\{\bar{\mathbf{c}}\}$ via cosine similarity:

$$\hat{c} = \arg\max_{c} \mathrm{sim}(\mathbf{c}_{\mathcal{Q}}, \bar{\mathbf{c}}^{c}), \quad \bar{\mathbf{c}}^{c} \in \{\bar{\mathbf{c}}\}.$$

The top-1 match provides an estimation of the input category $\hat{c}$.

At the intermediate level, each patch token $\mathbf{p}_{\mathcal{Q}}$ from the input image queries the semantic prototype set $\{\bar{s}\}^{\hat{c}}$ of class $\hat{c}$, retrieving its top-$K'$ nearest semantic prototypes $\{\bar{\mathbf{s}}^{\hat{c},j}\}_{j=1}^{K'}$, defined as:

$$\{\bar{\mathbf{s}}^{\hat{c},j}\}_{j=1}^{K'} = \arg\max_{\bar{\mathbf{s}}\in\{\bar{\mathbf{s}}\}^{\hat{c}}}{}^{K'} \mathrm{sim}(\mathbf{p}_{\mathcal{Q}}, \bar{\mathbf{s}}^{\hat{c},j}).$$

Finally, at the lowest level, $\mathbf{p}_{\mathcal{Q}}$ further queries the instance token set $\{\mathbf{t}\}^{\hat{c},j}$ associated with its matched semantic prototypes $\{\bar{\mathbf{s}}^{\hat{c},j}\}_{j=1}^{K'}$, retrieving the top-$K$ most similar instance tokens $\{\mathbf{t}^{\hat{c},j,k}\}_{k=1}^{K}$, defined as:

$$\{\mathbf{t}^{\hat{c},j,k}\}_{k=1}^{K} = \arg\max_{\mathbf{t}\in\{\mathbf{t}\}^{\hat{c},j}}{}^{K} \mathrm{sim}(\mathbf{p}_{\mathcal{Q}}, \mathbf{t}^{\hat{c},j,k}).$$

Among the retrieved $K'$ semantic prototypes, only the most relevant one is retained for each patch token. Consequently, by efficiently traversing all patch tokens $\{\mathbf{p}_{\mathcal{Q}}\}$ of the input image, we prepare a total of $H' \times W' \times 1$ semantic prototypes and $H' \times W' \times K$ template tokens, each represented as a $1 \times D$ feature vector.

### 3.4. Filtering-based Generation Reasoning

While the hierarchical retrieval flow effectively gathers template tokens, it also introduces hallucinatory noise from unreliable matches, spatial misalignment, and domain shifts. This noise blurs anomaly boundaries and obscures subtle defects. To address it, we reformulate the generation stage of RAG as a filtering-based generative reasoning process, employing a guided MoE filter that adaptively denoises and refines the anomaly cost volume, as shown in Fig. 2(b).

**Initial anomaly cost volume.** For each query token $\mathbf{p}_{\mathcal{Q}}^{(y,x)}$ at spatial coordinate $(y, x)$, we define patch-level anomaly cost with its retrieved instance patches $\mathbf{t}^{(y,x),k} \in \{\mathbf{t}^{(y,x),k}\}_{k=1}^{K}$ based on cosine similarity in Eqn. (1):

$$\mathcal{C}^{y,x,k} = 1 - \mathrm{sim}(\mathbf{p}_{\mathcal{Q}}^{(y,x)}, \mathbf{t}^{(y,x),k}),$$

where $\mathcal{C} \in \mathbb{R}^{H' \times W' \times K}$ is the resulting 3D anomaly cost volume, with $(y, x) \in \mathbb{R}^{H' \times W'}$ indicating spatial positions and $k \in \mathbb{R}^{K}$ indexing the matching candidates. Note that lower similarity values correspond to higher anomaly likelihoods.

By leveraging the hierarchical retrieval mechanism, RAID selects only a small set of highly relevant candidates for each query token, resulting in a compact and well-aligned cost volume. This design contrasts with CostFilter-AD [68], which constructs a global matching space, thereby significantly improving inference efficiency.

**Guided MoE filtering.** The guided MoE filter is designed as a two-stage architecture: the first stage constructs a guidance map via dual-guidance fusion, while the second stage performs guided filtering.

In the first stage, the semantic prototypes $\{\bar{\mathbf{s}}\}$ are rearranged into image-like prototype guidance maps $g_s \in \mathbb{R}^{H' \times W' \times D \times 1}$, and the query tokens $\{\mathbf{p}_{\mathcal{Q}}\}$ are rearranged into image-like query guidance maps $g_{\mathcal{Q}} \in \mathbb{R}^{H' \times W' \times D \times K}$. A convolutional router takes the concatenation $\mathrm{cat}(g_{\mathcal{Q}}, g_s)$ as input to compute sparse routing probabilities and weights, then aggregates the activated experts to produce the fused guidance $\tilde{g}$:

$$p = \mathrm{Softmax}(\mathrm{Router}(\mathrm{cat}(g_{\mathcal{Q}}, g_s))),$$

$$\tilde{p}^i = \begin{cases} p^i, & i \in \text{Top-}k(p), \\ 0, & \text{otherwise}, \end{cases}$$

$$\tilde{g} = \sum\nolimits_{i=1}^{M} \tilde{p}^i E_g^i \left(\mathrm{cat}(g_{\mathcal{Q}}, g_s)\right),$$

where $E_g^i$ denotes the $i$-th convolutional expert used for guidance fusion, and $\tilde{p}^i$ is its routing weight. Only the top-$k$ experts with the highest scores are activated, encouraging specialization across distinct semantic patterns.

In the second stage, the initial anomaly cost volume $\mathcal{C}$ is refined via denoising MoE, where a router softly activates all experts. Each denoising expert $E_{\mathcal{C}}^i$ performs dual-branch filtering of $\mathcal{C}$ under the fused guidance $\tilde{g}$, which consists of a cross-attention branch ($\tilde{g}$ as query and $\mathcal{C}$ as key/value) and a convolutional branch. The router assigns dense expert weights $p^i$ and aggregates the expert outputs $\tilde{\mathcal{C}}^i$ to yield the final anomaly map:

$$\mathcal{M} = \sum\nolimits_{i=1}^{M} p^i \cdot \tilde{\mathcal{C}}^i, \;\; \tilde{\mathcal{C}}^i = E_{\mathcal{C}}^i(\tilde{g}, \mathcal{C}),$$

where $\mathcal{M}$ denotes the generated anomaly map. The complete MoE architecture is detailed in the Appendix.

### 3.5. Training and Inference

We adopt a self-supervised training strategy commonly used in UAD [59, 62, 63], where synthetic anomalous images are paired with their corresponding synthetic anomaly masks $\mathcal{M}_s$. The overall objective function is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{focal}}(\mathcal{M}, \mathcal{M}_s) + \lambda_{\text{bal}} \mathcal{L}_{\text{bal}},$$

where the focal loss [35] $\mathcal{L}_{\text{focal}}$ addresses the inherent imbalance between normal and anomalous pixels, $\mathcal{L}_{\text{bal}}$ regularizes the expert routing process to prevent bias toward dominant experts and mitigate router collapse [14], and $\lambda_{\text{bal}}$ is the balancing weight.

For inference, the anomaly cost volume is constructed and filtered to yield the refined anomaly map $\mathcal{M}$. The image-level anomaly score is computed as the mean of the top 1% highest responses in $\mathcal{M}$, while for pixel-level localization, $\mathcal{M}$ is directly used as the anomaly map.

## 4. Experiments

### 4.1. Experimental Settings

**Datasets.** We evaluate our method on four widely used industrial anomaly detection benchmarks, MVTec-AD [4], VisA [75], MPDD [27], and BTAD [44], which cover diverse anomaly types, object complexities, and imaging conditions, offering a comprehensive benchmark for assessing UAD robustness and generalization. **MVTec-AD** contains 5,354 high-resolution images of 10 objects and 5 textures, with normal samples for training and diverse defects for testing. **VisA** includes 10,821 images across 12 object subsets, covering various surface and structural anomalies such as scratches, dents, and color spots. **MPDD** focuses on metallic parts, providing 1,346 images under challenging backgrounds and illumination. **BTAD** comprises 2,830 images from three industrial categories, featuring both normal and defective samples with real-world variability.

**Evaluation metrics.** Following common practice, we primarily report the Area Under the Receiver Operating Characteristic Curve for image-level anomaly detection (I-AUROC) and pixel-level anomaly localization (P-AUROC). Additional metrics, including Average Precision (AP), the maximum F1 score (F1-max), and the pixel-level Area Under the Per-Region Overlap (AUPRO), are also evaluated.

**Implementation details.** For full-shot and multi-dataset experiments, all input images are resized to $256 \times 256$. We use DINOv2-s [46] as the feature extractor. In constructing the hierarchical vector database, 80 templates are used for MVTec-AD and 100 for VisA, while all normal samples serve as templates for MPDD and BTAD. Patch tokens within each class are clustered into 50 semantic prototypes. During retrieval, the $K' = 5$ nearest semantic prototypes and $K = 150$ instance tokens are retrieved for each query

Table 1. **Full-shot** (multi-class UAD) performance (%) on four industrial datasets. Best results are highlighted in **bold**.

| Dataset | Method | Image-level | | | Pixel-level | | | |
|---|---|---|---|---|---|---|---|---|
| | | AUROC | AP | F1-max | AUROC | AP | F1-max | AUPRO |
| MVTec-AD [4] | PatchCore [49] | 96.4 | - | - | 95.7 | - | - | - |
| | UniAD [60] | 96.5 | 98.8 | 96.2 | 96.8 | 43.4 | 49.5 | 90.7 |
| | SimpleNet [39] | 95.3 | 98.4 | 95.8 | 96.9 | 45.9 | 49.7 | 86.5 |
| | MambaAD [22] | 98.6 | 99.6 | 97.8 | 97.7 | 56.3 | 59.2 | 93.1 |
| | GLAD [59] | 97.5 | 98.8 | 96.8 | 97.3 | 58.8 | 59.7 | 92.8 |
| | DiAD [23] | 97.2 | 99.0 | 96.5 | 96.8 | 52.6 | 55.5 | 90.7 |
| | ViTAD [64] | 98.3 | 99.4 | 97.3 | 97.7 | 55.3 | 58.7 | 91.4 |
| | AnomalyDINO [11] | 96.8 | 98.6 | 97.1 | 98.1 | 61.3 | 60.8 | 93.6 |
| | Costfilter-AD [68] | 99.0 | 99.7 | 98.6 | 98.0 | 58.1 | 61.2 | 93.2 |
| | RAID (Ours) | **99.4** | **99.8** | **98.7** | **98.6** | **71.7** | **68.5** | **95.6** |
| VisA [75] | UniAD [60] | 85.5 | 85.5 | 84.4 | 95.9 | 21.0 | 27.0 | 75.6 |
| | SimpleNet [39] | 87.2 | 87.0 | 81.8 | 96.8 | 34.7 | 37.8 | 81.4 |
| | MambaAD [22] | 94.3 | 94.5 | 89.4 | 98.5 | 39.4 | 44.0 | 91.0 |
| | GLAD [59] | 90.1 | 91.4 | 86.7 | 97.4 | 33.9 | 39.4 | 91.5 |
| | DiAD [23] | 86.8 | 88.3 | 85.1 | 96.0 | 26.1 | 33.0 | 75.2 |
| | ViTAD [64] | 90.5 | 91.7 | 86.3 | 98.2 | 36.6 | 41.1 | 85.1 |
| | AnomalyDINO [11] | 90.5 | 91.4 | 86.2 | 97.5 | 39.6 | 40.4 | 86.3 |
| | Costfilter-AD [68] | 93.4 | 95.2 | 89.3 | 98.6 | 41.4 | 45.0 | 86.8 |
| | RAID (Ours) | **94.9** | **95.5** | **90.6** | **99.0** | **45.2** | **49.2** | **91.7** |
| MPDD [27] | PatchCore [49] | 83.5 | - | - | 97.7 | - | - | - |
| | UniAD [60] | 80.1 | 83.2 | 85.1 | 95.4 | 19.0 | 25.6 | 83.8 |
| | Hvq-Trans [60] | 86.5 | 87.9 | 85.6 | 96.9 | 26.4 | 30.5 | 88.0 |
| | SimpleNet [39] | 90.6 | 94.1 | 89.7 | 97.1 | 33.6 | 35.7 | 90.0 |
| | MambaAD [22] | 89.2 | 93.1 | 90.3 | 97.7 | 33.5 | 38.6 | 92.8 |
| | GLAD [59] | 90.8 | 90.5 | 90.2 | 98.0 | 40.0 | 40.6 | 93.1 |
| | DiAD [23] | 85.8 | 89.2 | 86.5 | 91.4 | 15.3 | 19.2 | 66.1 |
| | ViTAD [64] | 87.4 | 90.8 | 87.0 | 97.8 | 44.1 | 46.4 | 95.3 |
| | Costfilter-AD [68] | 93.1 | 95.4 | 90.3 | 97.5 | 34.1 | 37.0 | 82.9 |
| | RAID (Ours) | **96.3** | **97.6** | **95.0** | **98.9** | **47.0** | **46.9** | **96.6** |
| BTAD [44] | UniAD [60] | 94.5 | 98.4 | 94.9 | 97.4 | 52.4 | 55.5 | **78.9** |
| | Hvq-Trans [60] | 90.9 | 97.8 | 94.8 | 96.7 | 43.2 | 48.7 | 75.6 |
| | SimpleNet [39] | 94.0 | 97.9 | 93.9 | 96.2 | 41.0 | 43.7 | 69.6 |
| | MambaAD [22] | 92.9 | 96.2 | 93.0 | **97.6** | 51.2 | 55.1 | 77.3 |
| | DiAD [23] | 90.2 | 88.3 | 92.6 | 91.9 | 20.5 | 27.0 | 70.3 |
| | ViTAD [64] | 94.0 | 97.0 | 93.7 | **97.6** | 58.3 | 56.5 | 72.8 |
| | Costfilter-AD [68] | 93.3 | **98.6** | **96.0** | 97.3 | 47.0 | 50.2 | 76.2 |
| | RAID (Ours) | **95.2** | 96.3 | 93.0 | **97.6** | **67.3** | **64.3** | 72.2 |

token. The guided MoE filter in the generation stage includes three experts per layer, with sparse routing in the first layer activating two experts per input. The weighting parameter $\lambda_{\text{bal}}$ is 0.005. Models are trained for 100 epochs using Adam with an initial learning rate of $1 \times 10^{-4}$. For few-shot evaluation, the same settings are applied except that input images are resized to $224 \times 224$ and the number of templates is adjusted accordingly.

## 4.2. Quantitative Comparison

We comprehensively evaluate the effectiveness and generalization of RAID under the multi-class UAD paradigm. The extensive experiments are conducted on four widely used benchmarks: MVTec-AD [4], VisA [75], MPDD [27], and BTAD [44], under three representative settings: **full-shot**, **few-shot**, and **multi-dataset**. For the **full-shot** scenario, we adopt the "one-model-for-all classes" paradigm and compare our method with a diverse set of State-Of-The-Art (SOTA) baselines, including PatchCore [49], HVQ-Trans [41], GLAD [59], DiAD [23], ViTAD [64], CostFilter-AD [68], UniAD [60], SimpleNet [39], MambaAD [22], and AnomalyDINO (full-shot) [11], covering feature-embedding, diffusion-based, and transformer-based frameworks. For the **few-shot** scenario, we evaluate against PatchCore [49], Win-CLIP [26], FastRecon [13], PromptAD [33], IIPAD [42], and DFM [57], to assess the model performance using limited normal samples **without any fine-tuning**. For the **multi-dataset** scenario, we adopt the "one-model-for-all datasets" paradigm, jointly training on multiple datasets and comparing it with the representative OneNIP [16] to validate cross-dataset generalization.

**Full-shot for multi-class UAD.** We evaluate image-level detection and pixel-level localization for full-shot UAD, where the training phase has access to all normal samples on four benchmarks, as shown in Table 1. On **MVTec-AD** [4], our method *outperforms* GLAD [59], AnomalyDINO [11], and CostFilter-AD [68] by $1.9\%/1.3\%$, $1.6\%/0.5\%$, and $0.4\%/0.6\%$ in I-AUROC/P-AUROC, respectively. On **VisA** [75], it reaches SOTA with gains of $4.8\%/1.6\%$, $4.4\%/1.5\%$, and $1.5\%/0.4\%$. On **MPDD** [27], advances are $5.5\%$, $8.9\%$, and $3.2\%$ for I-AUROC and $0.9\%$, $1.1\%$, and $1.4\%$ for P-AUROC over GLAD, ViTAD, and CostFilter-AD. On **BTAD** [44], it exceeds CostFilter-AD and ViTAD by $1.9\%$ and $1.2\%$ in I-AUROC.

Table 2. **Few-shot** performance comparison under the **multi-class** paradigm on MVTec-AD and VisA using I-AUROC/P-AUROC.

| Setup | Datasets | PatchCore [49] | Win-CLIP [26] | FastRecon [13] | PromptAD [33] | IIPAD [42] | DFM [57] | RAID (Ours) |
|---|---|---|---|---|---|---|---|---|
| 1-shot | MVTec-AD | 86.3 / 93.3 | 92.6 / 91.6 | 83.7 / 93.9 | 93.0 / 95.2 | 94.2 / 96.4 | 87.2 / 95.2 | **95.1 / 96.6** |
| | VisA | 79.9 / 95.4 | 84.8 / 95.3 | 80.1 / 96.5 | 85.2 / 97.2 | 85.4 / 96.9 | 84.0 / 96.4 | **85.8 / 97.7** |
| 2-shot | MVTec-AD | 83.4 / 92.0 | 93.8 / 91.9 | 88.9 / 95.3 | 95.4 / 95.6 | 95.7 / 96.7 | 90.2 / 96.0 | **96.6 / 97.1** |
| | VisA | 81.6 / 96.1 | 83.5 / 95.7 | 84.6 / 97.5 | 85.1 / 97.7 | 86.7 / 97.2 | 86.0 / 96.8 | **88.5 / 97.9** |
| 4-shot | MVTec-AD | 88.8 / 94.3 | 95.5 / 92.4 | 94.2 / 95.9 | 95.9 / 96.0 | 96.1 / 97.0 | 92.0 / 96.2 | **96.9 / 96.9** |
| | VisA | 85.3 / 96.8 | 85.7 / 96.0 | 68.5 / 96.0 | 87.5 / 97.9 | 88.3 / 97.4 | **89.8** / 97.1 | 89.3 / **98.2** |

Table 3. **Multi-dataset** performance comparison of **a single model** jointly trained on MVTec-AD, VisA, MPDD, and BTAD using I-AUROC/I-AP/P-AUROC/P-AP.

| Datasets | # class | OneNIP [16] | RAID (Ours) |
|---|---|---|---|
| MVTec-AD | 15 | 96.7 / 98.8 / 97.1 / 57.5 | 99.4 / 99.8 / 98.5 / 69.4 |
| VisA | 12 | 92.8 / 95.1 / 98.7 / 41.1 | 94.6 / 94.8 / 98.8 / 47.1 |
| MPDD | 6 | 85.2 / 88.4 / 97.8 / 37.0 | 93.6 / 96.1 / 98.8 / 44.9 |
| BTAD | 3 | 93.2 / 96.5 / 97.9 / 59.8 | 93.8 / 96.0 / 97.8 / 66.4 |
| All | 36 | 92.0 / 94.7 / 97.9 / 48.9 | **95.4 / 96.7 / 98.5 / 57.0** |

In addition, our method also delivers strong performance on other metrics involving image-level AP/F1-max, and pixel-level AP/F1-max/AUPRO. Collectively, the consistent gains across datasets support our RAG-inspired design: hierarchical retrieval supplies contextually relevant references, and the guided filter refines the anomaly cost volume into reliable image-level decisions and precise pixel-level localization.

**Few-shot UAD generalizability.** We evaluate RAID in the few-shot setting, resizing inputs to $224 \times 224$ for a fair comparison with DFM [42], IIPAD [57], and PromptAD [57]. Following common practice, the model is trained on the auxiliary MPDD dataset and then tested on MVTec-AD [4] and VisA [75]. Table 2 reports averaged results over five seeds, RAID attains consistently high I-AUROC/P-AUROC (standard deviations in the Appendix). On MVTec-AD, it surpasses DFM [57] by $7.9\%/1.4\%$, $6.4\%/1.1\%$, and $4.9\%/0.7\%$ for the 1-, 2-, and 4-shot settings (I-AUROC/P-AUROC). Notably, RAID operates in an image-only manner, unlike PromptAD [33] and WinCLIP [26] that leverage language priors, and it uses a lower input resolution than FastRecon [13] and WinCLIP [26]. These results highlight the strong generalization capability of RAID, which redefines UAD through the RAG perspective: contextually relevant references retrieved from the database iteratively refine the anomaly cost volume and guide generation toward semantically coherent regions, leading to more accurate and transferable anomaly localization.

**Multi-dataset UAD scalability.** We merge MVTec-AD [4], VisA [75], BTAD [44], and MPDD [27] into a larger and more diverse dataset following OneNIP [16], and train RAID on this merged corpus. Table 3 reports image-level classification and pixel-level segmentation, including the overall average across 36 categories and the per dataset averages. In this multi-dataset setting, RAID surpasses OneNIP [16] in image-level classification $(95.4\%/96.7\%)$ and pixel-level segmentation $(98.5\%/57.0\%)$, whereas OneNIP attains $92.0\%/94.7\%/97.9\%/48.9\%$ on the corresponding metrics. Moreover, moving from the multi-class setting in Table 1 to the unified regime of the dataset across Table 3, our method shows no noticeable degradation, thus establishing the proposed RAID as a scalable and effective solution for UAD in complex distributions.

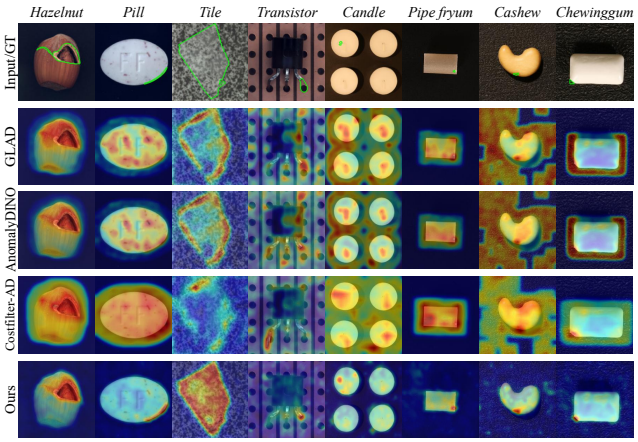


Figure 3. Qualitative comparison of multi-class anomaly localization results on MVTec-AD and VisA datasets.

## 4.3. Qualitative Comparison

We present qualitative comparisons on the MVTec-AD [4] and VisA [75] benchmarks to further validate the effectiveness of our method. As shown in Fig. 3, our approach outperforms GLAD [59], AnomalyDINO [11], and CostFilter-AD [68], achieving more precise anomaly localization with sharper boundaries, reduced matching noise, and enhanced sensitivity to subtle defects, owing to the RAG-like paradigm. Please refer to the Appendix for more visualization results.

## 4.4. Ablation Studies and Further Analysis

**Effectiveness of the retrieval strategy.** In Table 4, we compare two retrieval schemes (flat vs. hierarchical) under three database capacity settings (multi-class, single-class, and few-shot). In practice, the hierar-

Table 4. The efficiency and accuracy (I-AUROC and P-AUROC) comparison on MVTec-AD under the different retrieval and database (DB.) capacity settings.

| Methods | Retrieval | DB. | FLOPs | Mem. (GB) | Inf. (s) | Results |
|---|---|---|---|---|---|---|
| PatchCore | Flat | Multi | 7.12G | 3.46 | 0.093 | 96.4 / 95.7 |
| PatchCore | Flat | Single | 7.12G | 2.48 | 0.054 | 99.0 / 98.0 |
| AnomalyDINO | Flat | Single | 4.90G | 3.25 | 0.067 | 96.8 / 98.1 |
| RAID | Flat | Multi | 14.2G | 6.55 | 0.267 | 99.4 / 98.7 |
| RAID | Hier. | Single | 14.2G | 6.52 | 0.052 | 99.3 / 98.7 |
| RAID | Hier. | Multi | 14.2G | 6.52 | 0.062 | 99.4 / 98.6 |
| RAID | Hier. | 4-shot | 14.2G | 5.28 | 0.046 | 96.9 / 96.9 |

Table 5. Effectiveness of template quantity on MVTec-AD.

| # template | 20 | 40 | 60 | 80 | All |
|---|---|---|---|---|---|
| Metrics | 98.1 / 97.7 | 98.7 / 97.8 | 98.9 / 98.1 | **99.4 / 98.6** | 99.3 / **98.6** |

chical retrieval achieves about $5\times$ lower per-image latency than the flat scheme $0.267s$ (which prioritizes accuracy over speed) while maintaining nearly identical I-AUROC and P-AUROC of $99.4\%/98.6\%$. This demonstrates the efficiency–precision balance of our hierarchical retrieval scheme. Across the database capacities, multi-class and single-class settings yield comparable accuracy ($99.4\%/98.6\%$ vs. $99.3\%/98.7\%$), and even a 4-shot setup preserves strong performance $96.9\%/96.9\%$ at a similar per-image runtime ($0.046s$), highlighting the efficient use of limited normal templates. Overall, these observations validate the design of the hierarchical vector database and the associated retrieval strategy.

**Analysis of template quantity.** We vary the per-class template count on MVTec-AD from 20 to *All*, as shown in Table 5. The performance steadily improves from $98.1\%/97.7\%$ (I-AUROC/P-AUROC) to a peak of $99.4\%/98.6\%$ at 80 templates, while the *All* setting remains near-saturated at $99.3\%/98.6\%$. The gain up to 80 templates validates the retrieval pipeline: expanding the relevant template pool broadens normal-mode coverage and enhances query–template contrast, while hierarchical retrieval compacts matching candidates into a cleaner cost volume for the guided MoE to refine. The near-saturation under *All* suggests that, while further increasing the template count may offer marginal gains, relevance remains the dominant factor in achieving strong detection and localization with a compact, well-curated template database.

**Effectiveness of the guided MoE filter.** Table 6 analyzes how the guided MoE filter transforms the retrieval-built cost volume into reliable anomaly decisions. Starting from retrieval only (ID0, $97.9\%/97.5\%$), adding the cross-attention branch (Cro-Att.) and the Router$_{\mathcal{C}}$ in the 2nd MoE stage (ID1) raises accuracy to $98.5\%/97.6\%$ by refining semantically consistent matches. Incorporating the first-stage MoE$_g$ (ID2) further boosts accuracy to $99.2\%/98.4\%$, demonstrating the benefits of dual-guidance fusion. The convolutional branch (Conv.) in the second-stage (ID3) refines local responses ($98.7\%/97.8\%$), while its combination with cross-attention (ID4, $99.1\%/98.1\%$) or MoE$_g$ (ID5, $98.9\%/98.2\%$) brings complementary gains. The former strengthens denoising with a global perception, and the latter refines the guidance. Removing Router$_{\mathcal{C}}$ (ID6) leads to a drop ($98.0\%/97.5\%$), confirming the need for sparse routing to preserve expert specialization. The full configuration (ID7) achieves the best ($99.4\%/98.6\%$), validating that the guided MoE filter jointly denoises and refines the cost volume, delivering consistent improvements in both detection and localization.

Table 6. Ablation studies of guided MoE filter on MVTec-AD.

| ID | MoE$_g$ | Cro-Att. | Conv. | Router$_{\mathcal{C}}$ | I-/P-AUROC |
|---|---|---|---|---|---|
| 0 | - | - | - | - | 97.9 / 97.5 |
| 1 | - | ✓ | - | ✓ | 98.5 / 97.6 |
| 2 | ✓ | ✓ | - | ✓ | 99.2 / 98.4 |
| 3 | - | - | ✓ | ✓ | 98.7 / 97.8 |
| 4 | - | ✓ | ✓ | ✓ | 99.1 / 98.1 |
| 5 | ✓ | - | ✓ | ✓ | 98.9 / 98.2 |
| 6 | - | ✓ | ✓ | - | 98.0 / 97.5 |
| 7 | ✓ | ✓ | ✓ | ✓ | **99.4 / 98.6** |

Table 7. Effectiveness of expert quantity on MVTec-AD.

| # $(E_g, E_{\mathcal{C}})$ | (3, 1) | (3, 2) | (2, 3) | (3, 3) | (4, 3) |
|---|---|---|---|---|---|
| Metrics | 99.0 / 98.2 | 99.2 / 98.4 | 99.0 / 98.1 | **99.4 / 98.6** | 98.8 / 97.9 |

**Expert quantity in MoE.** Table 7 analyzes the impact of expert numbers in the two-stage guided MoE filtering $(E_g, E_{\mathcal{C}})$. Performance peaks at $(3, 3)$ with $99.4\%/98.6\%$ (I-AUROC/P-AUROC), indicating optimal specialization. Increasing the denoising experts $E_{\mathcal{C}}$ from 1 to 3 improves accuracy, while reducing guidance diversity $E_g = 2$ or over-expanding experts $E_g = 4$ degrades results due to reduced specialization and diluted guidance. These results highlight a capacity-specialization tradeoff, with $(3, 3)$ achieving the best balance for reliable anomaly detection.

## 5. Conclusion

We presented RAID, a Retrieval-Augmented Industrial anomaly Detection framework that revisits UAD from a RAG perspective. By integrating hierarchical retrieval with guided MoE filtering-based generation, RAID effectively suppresses matching noise and preserves fine-grained anomaly boundaries and subtle anomalies. Extensive experiments across full-shot, few-shot, and multi-dataset settings on four benchmarks demonstrate that RAID consistently outperforms prior SOTA methods, achieving both robust generalization and scalability. Looking forward, we envision that bringing the RAG paradigm into agentic and cross-modal anomaly detection opens a new direction toward more explainable, scalable, and data-efficient industrial intelligence.

## Acknowledgements

This work is supported in part by the Natural Science Foundation of Liaoning Province of China under Grant No. 2024-MSBA-42, by the Science and Technology Major Project of Liaoning Province under Grant No. 2024JH1/11700048, by the UKRI-AHRC CoSTAR National Lab for Creative Industries Research and Development No. AH/Y001060/1, and by the Key Research and Development Program of Liaoning Province under Grant No. 2023JH26/10200011.

## References

[1] Mouin Ben Ammar, Arturo Mendoza, Nacim Belkhir, Antoine Manzanera, and Gianni Franchi. Foundation models and transformers for anomaly detection: A survey. *Information Fusion*, 126:103517, 2026. 1

[2] Muhammad Aqeel, Shakiba Sharifi, Marco Cristani, and Francesco Setti. Towards real unsupervised anomaly detection via confident meta-learning. In *Proceedings of the ieee/cvf international conference on computer vision*, pages 4858–4867, 2025. 1

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

[4] 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. 5, 6, 7

[5] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. In *Advances in Neural Information Processing Systems*, pages 15309–15324. Curran Associates, Inc., 2022. 2

[6] Yunkang Cao, Xiaohao Xu, Jiangning Zhang, Yuqi Cheng, Xiaonan Huang, Guansong Pang, and Weiming Shen. A survey on visual anomaly detection: Challenge, approach, and prospect. *arXiv preprint arXiv:2401.16402*, 2024. 2

[7] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024. 2

[8] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 3

[9] Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 1&2: 1st place on zero-shot ad and 4th place on few-shot ad. *arXiv preprint arXiv:2305.17382*, 2023. 2

[10] Zhewei Dai, Shilei Zeng, Haotian Liu, Xurui Li, Feng Xue, and Yu Zhou. Seas: few-shot industrial anomaly image generation with separation and sharing fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23135–23144, 2025. 3

[11] Simon Damm, Mike Laszkiewicz, Johannes Lederer, and Asja Fischer. Anomalydino: Boosting patch-based few-shot anomaly detection with dinov2. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1319–1329. IEEE, 2025. 1, 2, 3, 4, 6, 7

[12] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[13] Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, and Jimin Xiao. Fastrecon: Few-shot industrial anomaly detection via fast feature reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17481–17490, 2023. 6, 7

[14] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. 5

[15] Matic Fučka, Vitjan Zavrtanik, and Danijel Skočaj. Transfusion–a transparency-based diffusion model for anomaly detection. In *European conference on computer vision*, pages 91–108. Springer, 2024. 1

[16] Bin-Bin Gao. Learning to detect multi-class anomalies with just one normal image prompt. In *European Conference on Computer Vision*, pages 454–470. Springer, 2024. 6, 7

[17] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023. 2, 3, 4

[18] Hewei Guo, Liping Ren, Jingjing Fu, Yuwang Wang, Zhizheng Zhang, Cuiling Lan, Haoqian Wang, and Xinwen Hou. Template-guided hierarchical feature restoration for anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6447–6458, 2023. 1, 2, 3, 4

[19] Hang Guo, Tao Dai, Zhihao Ouyang, Taolin Zhang, Yaohua Zha, Bin Chen, and Shu-tao Xia. Refir: Grounding large restoration models with retrieval augmentation. *Advances in Neural Information Processing Systems*, 37:46593–46621, 2024. 3

[20] Jia Guo, Shuai Lu, Lei Fan, Zelin Li, Donglin Di, Yang Song, Weihang Zhang, Wenbing Zhu, Hong Yan, Fang Chen, et al. One dinomaly2 detect them all: A unified framework for full-spectrum unsupervised anomaly detection. *arXiv preprint arXiv:2510.17611*, 2025. 2

[21] 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 Computer Vision and Pattern Recognition Conference*, pages 20405–20415, 2025. 1, 2

[22] 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. *Advances in Neural Information Processing Systems*, 37:71162–71187, 2024. 2, 6

[23] 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*, pages 8472–8480, 2024. 2, 6

[24] Jinlei Hou, Yingying Zhang, Qiaoyong Zhong, Di Xie, Shiliang Pu, and Hong Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8791–8800, 2021. 1

[25] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11375–11385, 2024. 1

[26] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023. 1, 3, 6, 7

[27] Stepan Jezek, Martin Jonak, Radim Burget, Pavel Dvorak, and Milos Skotak. Deep learning-based defect detection of metal parts: evaluating current methods in complex conditions. In *2021 13th International congress on ultra modern telecommunications and control systems and workshops*, pages 66–71, 2021. 5, 6, 7

[28] Jooyeon Kim, Eulrang Cho, Sehyung Kim, and Hyunwoo J Kim. Retrieval-augmented open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17427–17436, 2024. 2, 3

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2

[30] Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18035–18043, 2025. 3

[31] Byeonghun Lee, Hyunmin Cho, Hong Gyu Choi, Soo Min Kang, Iljun Ahn, and Kyong Hwan Jin. Reference-based super-resolution via image-based retrieval-augmented generation diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10764–10774, 2025. 3

[32] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020. 3

[33] Yiting Li, Adam Goodge David, Fayao Liu, and Chuan-Sheng Foo. Promptad: Zero-shot anomaly detection using text prompts. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1082–1091, 2024. 6, 7

[34] Yuxin Li, Yaoxuan Feng, Bo Chen, Wenchao Chen, Yubiao Wang, Xinyue Hu, Baolin Sun, Chunhui Qu, and Mingyuan Zhou. Vague prototype-oriented diffusion model for multi-class anomaly detection. In *Forty-first International Conference on Machine Learning*, 2024. 2

[35] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. 5

[36] 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*, page 103139, 2025. 1, 2, 3

[37] Heng Liu, Siru Jiang, Fangyun Duan, Yongzhe Lyu, Xiusong Wang, Hanlin Ge, and Chao Liang. Cadencerag: Context-aware and dependency-enhanced retrieval augmented generation for holistic video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3679–3688, 2025. 3

[38] Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Diversity-measurable anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12147–12156, 2023. 1

[39] 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*, pages 20402–20411, 2023. 2, 6

[40] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton Van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6959–6969, 2022. 2

[41] Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 36:8487–8500, 2023. 1, 2, 6

[42] 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*, 2025. 3, 6, 7

[43] Shiyuan Meng, Wenchao Meng, Qihang Zhou, Shizhong Li, Weiye Hou, and Shibo He. Moead: A parameter-efficient model for multi-class anomaly detection. In *European Conference on Computer Vision*, pages 345–361. Springer, 2024. 2

[44] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics*, pages 01–06, 2021. 5, 6, 7

[45] 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 Computer Vision and Pattern Recognition Conference*, pages 20383–20394, 2025. 1

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2, 5

[47] Zhen Qu, Xian Tao, Xinyi Gong, ShiChen Qu, Xiaopei Zhang, Xingang Wang, Fei Shen, Zhengtao Zhang, Mukesh Prasad, and Guiguang Ding. Dictas: A framework for class-generalizable few-shot anomaly segmentation via dictionary lookup. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20519–20528, 2025. 1, 2, 3

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[49] 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*, pages 14318–14328, 2022. 1, 2, 3, 4, 6, 7

[50] Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, pages 21999–22027, 2024. 3

[51] Lei Shen, Kang Zhao, Zhipeng Jin, Wen Tao, Yi Yang, Cong Han, Shuanglong Li, Zhongmin Cai, and Lin Liu. Retrieval-augmented image captioning and generation with entity concepts enhancement for baidu multimodal advertising. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4324–4328, 2025. 3

[52] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1, 2

[53] Ryota Tanaka, Taichi Iki, Taku Hasegawa, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. Vdocrag: Retrieval-augmented generation over visually-rich documents. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24827–24837, 2025. 3

[54] 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*, pages 24511–24520, 2023. 2

[55] Shun Wei, Jielin Jiang, and Xiaolong Xu. Uninet: A contrastive learning-guided unified framework with feature selection for anomaly detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9994–10003, 2025. 1, 2

[56] Gaochang Wu, Yapeng Zhang, Lan Deng, Jingxin Zhang, and Tianyou Chai. Cross-modal learning for anomaly detection in complex industrial process: Methodology and benchmark. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3):2632–2645, 2025. 1

[57] Sheng Wu, Yimi Wang, Xudong Liu, Yuguang Yang, Runqi Wang, Guodong Guo, David Doermann, and Baochang Zhang. Dfm: Differentiable feature matching for anomaly detection. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15224–15233, 2025. 6, 7

[58] Zhiwei Yang, Jing Liu, and Peng Wu. Text prompt with normality guidance for weakly supervised video anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18899–18908, 2024. 1

[59] Hang Yao, Ming Liu, Zhicun Yin, Zifei Yan, Xiaopeng Hong, and Wangmeng Zuo. Glad: Towards better reconstruction with global and local adaptive diffusion models for unsupervised anomaly detection. In *European Conference on Computer Vision*, pages 1–17. Springer, 2024. 1, 2, 5, 6, 7

[60] 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*, 35:4571–4584, 2022. 1, 2, 6

[61] Qinhan Yu, Zhiyou Xiao, Binghui Li, Zhengren Wang, Chong Chen, and Wentao Zhang. Mramg-bench: A comprehensive benchmark for advancing multimodal retrieval-augmented multimodal generation. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3616–3626, 2025. 3

[62] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draema discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021. 2, 5

[63] Hui Zhang, Zheng Wang, Dan Zeng, Zuxuan Wu, and Yu-Gang Jiang. Diffusionad: Norm-guided one-step denoising diffusion for anomaly detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1, 5

[64] Jiangning Zhang, Xuhai Chen, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, Ming-Hsuan Yang, and Dacheng Tao. Exploring plain vit reconstruction for multi-class unsupervised anomaly detection. *CVIU*, 2025. 1, 6

[65] Kai Zhang, Zekai Zhang, Xihe Sun, Jingmeng Nie, Qinghui Chen, Han Hao, Jianyuan Guo, and Jinglin Zhang. Adseeker: A knowledge-infused framework for anomaly detection and reasoning. *arXiv preprint arXiv:2508.03088*, 2025. 3

[66] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6782–6791, 2023. 1

[67] 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*, pages 3914–3923, 2023. 1, 2

[68] Zhe Zhang, Mingxiu Cai, Hanxiao Wang, Gaochang Wu, Tianyou Chai, and Xiatian Zhu. Costfilter-ad: Enhancing anomaly detection through matching cost filtering. In *International Conference on Machine Learning*, 2025. 2, 3, 4, 5, 6, 7

[69] Zhe Zhang, Mingxiu Cai, Gaochang Wu, Jing Zhang, Lingqiao Liu, Dacheng Tao, Tianyou Chai, and Xiatian Zhu. Unified unsupervised anomaly detection via matching cost filtering. *arXiv preprint arXiv:2510.03363*, 2025. 2

[70] Lin Zhao, Xiao Chen, Eric Z Chen, Yikang Liu, Terrence Chen, and Shanhui Sun. Retrieval-augmented few-shot medical image segmentation with foundation models. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. 3

[71] Ying Zhao. Omnial: A unified cnn framework for unsupervised anomaly localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3924–3933, 2023. 1, 2

[72] Xu Zheng, Ziqiao Weng, Yuanhuiyi Lyu, Lutao Jiang, Haiwei Xue, Bin Ren, Danda Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. Retrieval augmented generation and understanding in vision: A survey and new outlook. *arXiv preprint arXiv:2503.18016*, 2025. 2, 3, 4

[73] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023. 2, 3

[74] Yixuan Zhou, Xing Xu, Jingkuan Song, Fumin Shen, and Heng Tao Shen. Msflow: Multiscale flow-based framework for unsupervised anomaly detection. *IEEE transactions on neural networks and learning systems*, 2024. 2

[75] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pretraining for anomaly detection and segmentation. In *Computer Vision – ECCV 2022*, pages 392–408, Cham, 2022. Springer Nature Switzerland. 5, 6, 7