ReinAD: Towards Real-world Industrial Anomaly Detection with a Comprehensive Contrastive Dataset

Xu Wang[1,*], Jingyuan Zhuo[1,*], Zhiyuan You[2], Zhiyu Tan[1], Yikuan Yu[1], Siyu Wang[1], Xinyi Le[1,†]
[1]Shanghai Jiao Tong University, [2]The Chinese University of Hong Kong
{wx0413, mui123, tttangerine, yyyykkkk1995, y_wsy09, lexinyi}@sjtu.edu.cn
zhiyuanyou@foxmail.com

Abstract

Recent years have witnessed significant advancements in industrial anomaly detection (IAD) thanks to existing anomaly detection datasets. However, the large performance gap between these benchmarks and real industrial practice reveals critical limitations in existing datasets. We argue that the mismatch between current datasets and real industrial scenarios becomes the primary barrier to practical IAD deployment. To this end, we propose **ReinAD** dataset, a comprehensive contrastive dataset towards **Re**al-world **in**dustrial **A**nomaly **D**etection. Our dataset prioritizes three critical real-world requirements: 1) Contrast-based anomaly definition that is essential for industrial practice, 2) Fine-grained unaligned image pairs reflecting real inspections, and 3) Large-scale data from active production lines spanning multiple industrial categories. Based on our dataset, we introduce the ReinADNet. It takes both normal reference and test images as inputs, achieving anomaly detection through normal-anomaly comparison. To address the fine-grained and unaligned properties of real industrial scenes, our method integrates pyramidal similarity aggregation for comprehensive anomaly characterization and global-local feature fusion for spatial misalignment tolerance. Our method outperforms all baselines on the ReinAD dataset (*e.g.*, 64.5% *v.s.* 59.5% in 1-shot image-level AP) under all settings. Extensive experiments across several datasets demonstrate our dataset's challenging nature and our method's superior generalization. This work provides a solid foundation for practical industrial anomaly detection. Dataset and code are available at https://tocmac.github.io/ReinAD.

1 Introduction

Industrial anomaly detection (IAD) has made significant progress in recent years, benefiting from datasets such as MVTecAD [6], MPDD [33], BTAD [36], VisA [68], *etc.* Existing anomaly detection methods [32, 42, 57] have achieved remarkably high performance on these benchmarks. For example, PatchCore [42] has achieved an image-level AUROC higher than 99% on MVTecAD. However, these methods remain difficult to apply in real industrial scenarios [4, 37, 46, 49, 64]. This is mainly due to the gap between existing dataset and real industrial scenarios.

First, the contrastive ability is necessary for industrial anomaly detection. In real industrial scenarios, the identification of "which part is anomalous" should be initiated based on normal samples or rules. Notably, many industrial anomalies, such as "wire missing" in Fig. 1a, cannot be detected even by humans without the reference of normal samples. In contrast, many anomalies in existing datasets are defined only by common sense (*e.g.*, "capsule crack" in Fig. 1a), making them easier

[*]Contributed Equally.
[†]Corresponding Authors.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

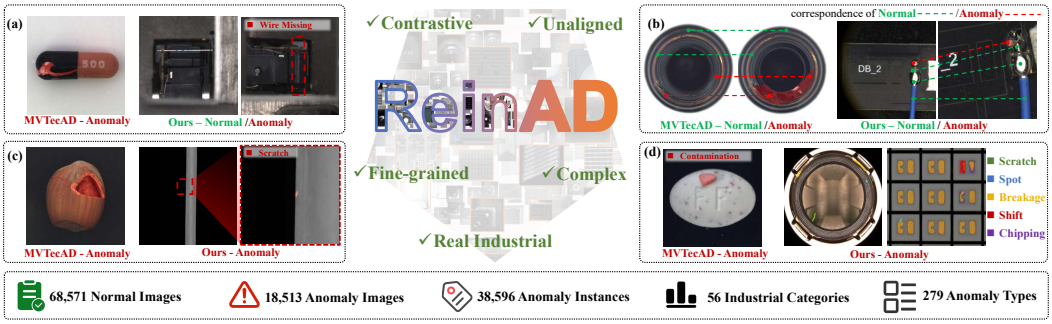


Figure 1: Illustration of our **ReinAD** dataset, a comprehensive contrastive dataset towards **Re**al-world **in**dustrial **A**nomaly **D**etection. (a) Some real anomalies (*e.g.*, "wire missing" circled by red) require contrast between normal and anomalous samples to detect. (b) Sample unalignment caused by variations in shifts, rotations, and scales in production environments. (c) Quite fine-grained anomalies (scratch) masked by red. (d) Multi-class anomalies may appear in one object.

to be identified even without normal references. This is evidenced by the fact that even 0-shot IAD method WinCLIP [32] has achieved a very high image-level AUROC (91.8%) on MVTecAD. Another advantage is that this contrastive ability can generalize to new categories *unseen* during training, as illustrated in prior works like InCTRL [67] and ResAD [55]. Therefore, we argue that the contrastive ability between normal and anomalous samples is crucial in industrial applications.

Second, the samples in existing datasets are mostly well aligned and the anomalies are obvious without complex categories, underestimating the challenges of actual industrial scenarios. As illustrated in Fig. 1b, all samples in the bottle category in MVTecAD dataset are well-aligned. However, images taken in real production lines exhibit variations in shifts, rotations, and scales (*e.g.*, right part in Fig. 1b) due to different production environments. Also, anomalies in existing datasets (*e.g.*, hazelnut crack in Fig. 1c) are usually obvious with a large size. In contrast, real anomalies shown in the right part of Fig. 1c can be extremely small and fine-grained. Finally, as depicted in Fig. 1d, multiple anomalies often co-occur on a single object, a scenario overlooked by existing datasets. Therefore, the difficulty of existing datasets is much lower than that of actual scenarios.

To address these challenges, we construct a large-scale dataset that matches better with real industrial demands, termed ReinAD. As illustrated in Fig. 1, Our dataset comprises four key components:

- **Contrastive capability.** We prioritize contrastive capability in sample collection. Many anomalies in our dataset can only be identified through comparison with normal samples.
- **Unaligned property.** Misalignment is common in real-world industrial imaging. Samples in our dataset capture this property through variations in shift, rotation, and scale.
- **Fine-grained anomalies.** Large quantities of anomalies in our dataset have a tiny area ratio, presenting significant challenges for anomaly detection.
- **Complex anomaly patterns.** Co-occurring anomalies are common in our dataset. This important real-world property is overlooked by many existing datasets.

Based on our ReinAD dataset, we propose ReinADNet, a model taking both normal reference and test image as inputs, identifying anomalies via comparing with normal reference. For fine-grained comparison, we propose a pyramidal cost aggregation module to compute point-wise multi-scale similarities. To contrast unaligned samples, we develop an adaptive nearest-neighbor search strategy for optimal local matching. Our method outperforms all baselines on ReinAD dataset (*e.g.*, 64.5% *v.s.* 59.5% in 1-shot image-level AP) under all settings. Cross-dataset experiments demonstrate both our dataset's challenges and our method's superior generalization.

In summary, our main contributions can be summarized as follows:

- We introduce ReinAD dataset, a novel dataset for real-world industrial anomaly detection. Our ReinAD dataset focuses on contrastive ability, containing unaligned samples and multi-class fine-grained anomalies, better reflecting real industrial scenes.
- Our comprehensive and large-scale ReinAD dataset provides a foundation for advanced anomaly detection methods. The introduced dataset contains 56 categories, 279 anomalous types, and 87,084 expert-annotated samples with anomalous segmentation masks.

- We propose ReinADNet, a generalizable anomaly detection method. ReinADNet identifies anomalies via normal-anomaly sample comparisons and handles fine-grained unaligned anomalies, achieving better results than previous baselines.

2 Related Works

Anomaly detection datasets. The evolution of anomaly detection datasets reflects incremental progress toward addressing real-world industrial challenges. Early works predominantly relied on KolektorSDD [44], a single-category dataset that constrained algorithm evaluation and development. Subsequent datasets like MTD [31], MPDD [33], and BTAD [36] expanded diversity but remained limited in scale and categorical coverage. A pivotal shift occurred with MVTec AD [6], standardizing industrial anomaly detection (IAD) research by providing 5,354 images across 15 object categories. VisA [68] further advanced this effort, scaling to 10,821 images spanning 12 objects and 15 anomaly types. However, existing datasets remain confined to narrow industrial scenarios due to their small scale and limited categories. Recent efforts like Real-IAD [46] introduced larger multi-view data, yet its reliance on artificially fabricated anomalies creates a significant domain gap in both object and anomaly realism. Meanwhile, domain-specific datasets (*e.g.*, VAD [3] for solder joints, CID [63] and CableInspect-AD [2] for cables, and 3CAD [52] for 3C components) focus on niche applications, limiting their utility for training models requiring generalizable anomaly detection capabilities across unseen industrial scenarios. These limitations underscore the urgent need for a large-scale, real-world dataset that captures the complexity and diversity of authentic industrial environments, enabling robust training and evaluation of models for generalizable anomaly detection.

Classical anomaly detection methods. Existing unsupervised anomaly detection methods exhibit three primary technical streams: 1) Distance-based approaches [17, 18, 24, 30, 42, 56] identify anomalies through statistical deviations in feature space; 2) Reconstruction-based methods [1, 12, 13, 28, 39, 51, 53, 54, 58, 60, 61] employ autoencoders or GANs to detect reconstruction errors; 3) Knowledge distillation-based methods [7, 9, 19, 43, 45, 47, 48] utilize teacher-student feature discrepancies. While achieving category-specific effectiveness, these methods inherently overfit to closed-set normal patterns and lack generalizable cross-category reasoning capabilities.

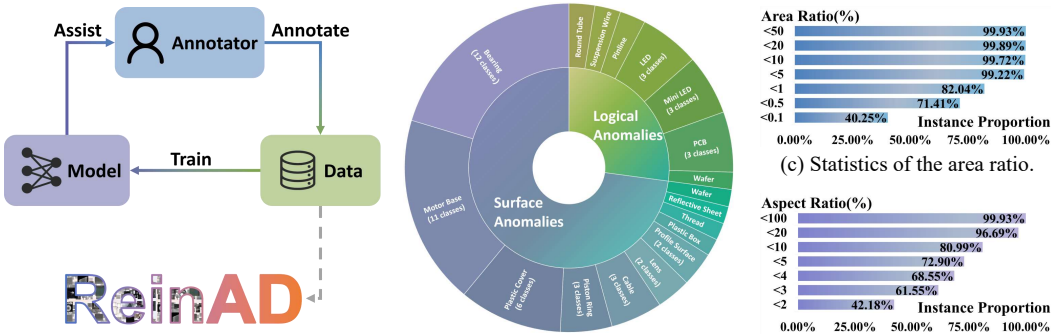
Prompt-based anomaly detection methods. Recent works leverage vision-language models (VLMs) like CLIP [41] for zero-shot detection [10, 14, 15, 23, 29, 32, 35, 40, 66], bypassing category-specific training via textual prompts. However, their performance depends critically on manual prompt design. Fixed templates show category inconsistency [11, 65], while dynamic prompts face semantic ambiguity in defining anomalies. Fundamentally, both classical and prompt-based methods focus on normality modeling rather than systematic anomaly reasoning, limiting their generalization capability.

Generalizable anomaly detection. Generalizable anomaly detection (GAD) seeks to develop unified detection models capable of generalizing across diverse application domains without requiring target-domain training data. The pioneering work InCTRL [67] established a baseline framework for cross-dataset anomaly classification by capturing contextual residuals between query images and normal references. While demonstrating category-generalizable detection capability, this method lacks precise anomaly localization, a critical requirement for industrial inspection scenarios. Subsequent work ResAD [55] addresses this limitation through residual feature learning with explicit normality constraints, enabling simultaneous detection and localization. Nevertheless, ResAD inherits fundamental constraints from traditional distance-based methods since its residual computation relies on global feature matching that ignores inter-image contextual relationships and intra-image neighborhood dependencies, thereby limiting its adaptability to complex anomaly patterns.

3 ReinAD Dataset

3.1 Dataset Construction

Data collection. Our data originates from multi-year accumulations in real industrial scenarios such as 3C electronics, mechanical components, consumer goods, *etc.* To address practical inspection needs, we develop customized optical solutions tailored for different workpieces and anomaly types (*e.g.* low-angle ring light for scratches and multi-zone light for dents), ensuring comprehensive coverage across diverse scenarios and production lines. Technicians then define anomaly criteria based on actual quality requirements and industrial SOP standards. During production, large quantities of both normal and anomaly samples are automatically captured, and subsequently labeled by annotators.



(a) Data annotation process. (b) Distribution of dataset categories. (c) Statistics of the area ratio. (d) Statistics of the aspect ratio.

Figure 2: **Data annotation process and statistics** of our ReinAD dataset. (a) Data annotation process. (b) Distribution of dataset categories. (c) Statistics of the anomaly area ratio of the anomaly images. (d) Statistics of the aspect ratio of the anomaly area's minimum bounding box.

Table 1: **Comparison with existing popular anomaly detection datasets**. "%AR<0.1" denotes the percentage of samples in which the ratio of anomalous area is less than 0.1%. Missing values (*i.e.*, "-") indicate data unavailable up to submission.

Dataset	Time	Class	Anomaly Types	Image Number			Anomaly Source	%AR<0.1
				Normal	Anomaly	Total		
KSDD [44]	2019	1	1	347	52	399	Real	8.47
MVTecAD [6]	2019	15	73	4,096	1,258	5,354	Human-crafted	1.01
MTD [31]	2020	1	6	952	392	1,344	Real	7.38
KSDD2 [8]	2021	1	5	2,979	356	3,335	Real	2.04
MPDD [33]	2021	6	–	1,064	282	1,346	Real	11.93
BTAD [36]	2021	3	9	2540	290	2,830	Real	6.09
VisA [68]	2022	12	75	9,621	1,200	10,821	Human-crafted	31.28
MIAD [5]	2023	7	14	87,500	17,500	105,000	Virtual	20.64
Real-IAD [46]	2024	30	131	99,721	51,329	151,050	Human-crafted	–
VAD [3]	2024	1	21	3,000	2,000	5,000	Real	–
CID [63]	2024	1	6	4,060	233	4,293	Real & Synthetic	–
CableInspect-AD [2]	2024	3	7	2,159	2,639	4,798	Real	0.92
3CAD [52]	2025	8	47	15,577	11,462	27,039	Real	28.65
MVTecAD-2 [26]	2025	8	20	4,705	3,299	8,004	Real	33.65
Ours	2025	56	279	68,571	18,513	87,084	Real	40.25

Data annotation. As illustrated in Fig. 2a, we design a human-in-the-loop semi-automated annotation pipeline. First, annotators manually label a small subset of samples according to predefined anomaly criteria. These annotated samples then serve as an initial training set for a segmentation model [25, 50]. The trained model subsequently generates preliminary annotations for the remaining unlabeled data. Next, human annotators refine these annotations to produce the final high-quality ground truth. Importantly, the newly annotated data are iteratively used to retrain and improve the model. This creates a positive feedback loop that progressively enhances the model's pre-annotation accuracy. Despite this optimized semi-automated approach, pixel-level annotations for our dataset remain labor-intensive. The entire annotation process requires about 600 person-hours of expert-level effort. Quantitative details on annotation quality improvement with the human-in-the-loop annotation pipeline are available in the supplementary material.

3.2 Dataset Description

Statistics. The statistics in Fig. 2b-d demonstrate the remarkable diversity of our dataset. Fig. 2b presents the category distribution of our dataset. The anomaly types can be broadly categorized into surface anomalies and logical anomalies. Our dataset encompasses 19 industrial categories, including daily necessities, 3C components, *etc.* Each category contains one or multiple distinct products. This diversity enhances our dataset's broad applicability. As shown in Fig. 2c, our dataset contains both large-scale and small-scale anomaly regions. Fig. 2d displays the aspect ratio distribution of anomaly areas' minimum bounding boxes, revealing diverse morphological characteristics of anomalies. Both the anomaly area proportions and aspect variations indicate our dataset's high difficulty level. This is

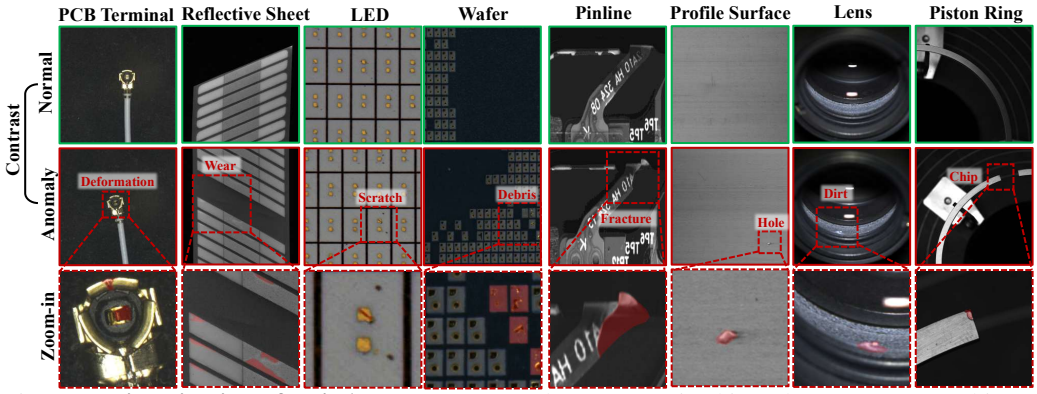


Figure 3: **Visualization of ReinAD dataset**. Samples are organized into three rows: normal images with green borders (top); anomaly images with red borders (middle); and Zoom-in patches of images in the middle row (bottom). The top black texts indicates object categories, and the red texts represents anomaly types. Additional visualizations are available in the supplementary material.

further corroborated by the experimental results in Tab. 2. We adapt a cross-category split between training and test sets to evaluate the model's generalization capability. The categories in training and test sets are completely distinct. They are randomly split while maintaining the same proportion of surface defects and logical defects.

Comparison with popular datasets. As shown in Tab. 1, our ReinAD exhibits four key advantages over existing anomaly detection datasets. First, our dataset contains a substantial number of fine-grained anomalies, reflecting the real challenges in industrial inspection scenarios. In our dataset, samples with anomaly area ratios below 0.1% account for over 40% of the total, surpassing all other datasets listed in the table. The second-highest ratio is only 33.65% for MVTecAD-2 [26], while popular datasets like MVTec AD [6], BTAD [36] and MPDD [33] show significantly lower proportions at just 1.01%, 6.09% and 11.93% respectively. Such subtle anomalies are actually common in real industrial settings, yet current datasets notably oversimplify this critical aspect. Second, our data are entirely sourced from real industrial scenarios. All anomalies in our dataset occurred naturally during manufacturing processes. This ensures authentic representation of industrial production. In contrast, widely used datasets such as MVTec AD [6], VisA [68], and Real-IAD [46] rely on human-crafted anomalies. Such artificial anomalies exhibit significant gaps compared to real-world cases. These gaps manifest in both anomaly feature granularity and diversity of anomaly types. Third, our ReinAD serves as a comprehensive industrial dataset, offering significantly more diverse object classes and anomaly types than existing datasets. Recent datasets, such as VAD [68], CID [63], CableInspect-AD [2], and 3CAD [52], focus on specific applications (*e.g.*, solder joints, cables, or 3C components). This limits their utility for training models requiring generalizable capabilities across unseen industrial scenarios. Fourth, our dataset surpasses most datasets (except MIAD [5] and Real-IAD [46]) in data scale. Notably, MIAD is a virtual simulation dataset, and anomalies of Real-IAD are human-crafted. To our best knowledge, our dataset represents the largest real-world industrial anomaly detection dataset.

Property analysis. As illustrated in Fig. 1 and Fig. 3, our dataset exhibits four key characteristics. 1) *Contrastive requirement:* Many anomalies in current datasets can be simply detected without normal reference. However, real industrial anomalies (*e.g.* the PCB terminal deformation in Fig. 3) can only be identified through comparison with normal samples. 2) *Unaligned property:* Real-world industrial imaging often involves imperfect alignment. Samples in our dataset capture this through variations in shift, rotation, and scale. In Fig. 3, the wear anomalies on the reflective sheet demonstrate this characteristic. 3) *Fine-grained Anomalies:* Our dataset contains subtle anomalies in industrial settings, exemplified by the LED scratch in Fig. 3. Quantitative analysis in Fig. 2c reveals that over 40% of anomalies in our dataset have a area ratio below 0.1%, presenting significant detection challenges. 4) *Complex anomaly patterns:* Co-occurring anomalies (*e.g.* multiple wafer debris anomalies in Fig. 3) are common in our dataset. This important property is overlooked by many current datasets.

4 ReinADNet Method

Problem statement. Our objective is to achieve fine-grained, general anomaly detection. Under the contrastive paradigm, the model must jointly learn normal and anomalous patterns and transfer

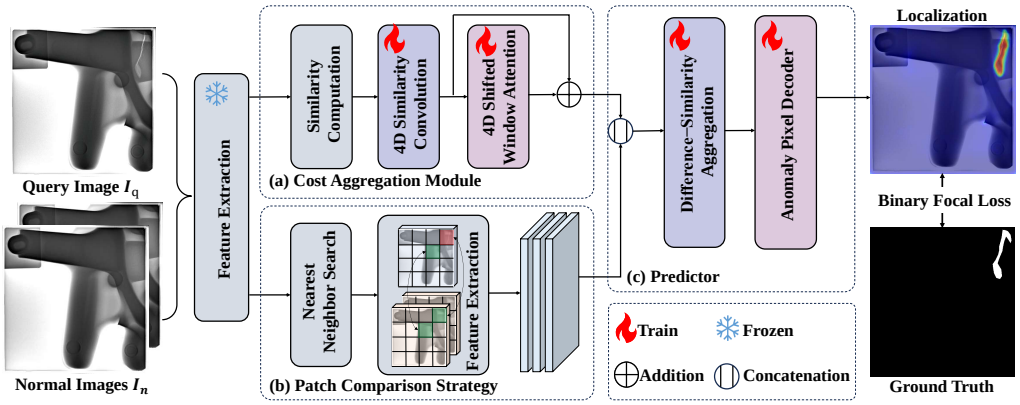


Figure 4: **Framework of our ReinADNet**. Given a query image and a set of reference images as input, a pretrained network extracts multi-scale features. The Cost Aggregation Module captures global point-to-point similarity between I_q and I_n, while the Patch Comparison Module captures local discrepancies. The predictor subsequently aggregates these discrepancy and similarity features to generate precise pixel-wise anomaly predictions. Additional information is available in the supplementary material.

this discriminative ability to novel categories. To emulate such scenarios, we use a source dataset D_{src} for training, where each subclass comprises normal samples I_n, anomalous samples I_q and corresponding masks M. In training, it randomly samples normal–anomalous or normal–normal pairs across all classes, supervised by the ground-truth masks. In testing, a structurally similar target dataset D_{tgt} containing unseen categories is used to evaluate the model's generalization.

Overview of our approach. As shown in Fig. 4, we extract multi-level features from the query image I_q and a set of normal images I_n, forming multiple feature pairs. The cost aggregation module computes and refines the global similarity for each pair, enabling fine-grained contrast. The patch comparison strategy uses prototype learning to detect anomalies at a local scale, effectively addressing the misalignment problem. Finally, the predictor combines similarity and difference cues and further integrates each pixel's neighborhood context to decode and output the anomaly heatmap.

Cost aggregation module. To address global semantic shifts, we adopt insights from relevant research [16, 27] on semantic matching tasks, enabling the model to directly learn feature-to-feature similarity. First, we compute multi-level similarity between query features f_q^l and normal features f_n^l across L hierarchical levels. Initial cost maps C^l are derived via cosine similarity, where i and j represent the 2D spatial positions of f_q^l and f_n^l:

$$C^l(i,j) = \frac{f_q^l(i) \cdot f_n^l(j)}{\|f_q^l(i)\| \|f_n^l(j)\|}. \tag{1}$$

Stacked cost maps $C \in \mathbb{R}^{h_q \times w_q \times h_n \times w_n \times L}$ undergo volumetric processing. A 4D CNN extracts multi-level features, followed by a 4D Swin Transformer for coarse-to-fine refinement:

$$M^l = \text{Conv4d}(C^l), \quad A^l = \text{Swin4d}(M^l), \quad A^{l-1} = \text{Swin4d}(M^l + \text{up}(A^l)). \tag{2}$$

Final feature $A \in \mathbb{R}^{h_q \times w_q \times C}$ is obtained.

Patch-level comparison strategy. Aligned with the paradigm of prototype learning, we propose a multi-scale patch comparison strategy. For each position (i,j) in query feature f_q, we compute cosine distances to all patches in normal feature f_n, identify the closest prototype f_{close}, and derive local discrepancy features f_{dist} as:

$$f_{close} = f_n \left(\arg\min \left(1 - \frac{f_q \cdot f_n}{\|f_q\| \|f_n\|} \right) \right), \quad f_{dist}(i,j) = f_{close}(i,j) - f_q(i,j). \tag{3}$$

Normal patches exhibit minimal f_{dist}, while anomalies yield larger mismatches. By deploying this module across multiple network layers, we capture scale-adaptive discrepancy features.

Predictor. A Swin Transformer-based module fuses the aggregated similarity features A with multi-scale discrepancy features f_{dist}:

$$f_{fussion} = \text{Swin2d}(A \oplus f_{dist}), \quad m = \text{Conv2d}(f_{fussion}), \tag{4}$$

where $\oplus$ denotes concatenation. Predictor integrates global-local context, and generates anomaly heatmaps, with maximum anomaly score as image-level output.

Training. The image encoder remains frozen. Using one normal sample per class as reference, we train with normal/anomaly query pairs. Focal loss addresses class imbalance:

$$\mathcal{L} = \frac{1}{N} \sum_{x \in D_{src}} \mathcal{L}(S(x), G(x)), \tag{5}$$

where $S(x)$ is the predicted heatmap and $G(x)$ the ground truth.

Inference. For a test image, we compare it against reference normal samples to generate a patch-level heatmap. The maximum heatmap value determines the image-level anomaly score.

5 Experiments

5.1 Experimental Setup

Datasets and metrics. To assess both our dataset's challenge and our method's generalization capability, we conduct comprehensive experiments across our ReinAD dataset and several popular datasets. These datasets include MVTecAD [6], VisA [68], BTAD [36], and MPDD [33]. Previous works typically rely solely on AUROC (Area Under the Receiver Operating Characteristic Curve) as an evaluation metric. However, in anomaly detection tasks, there exists a significant class imbalance between anomalous and normal pixels, with anomalous regions accounting for only a small fraction of the total. Consequently, AUROC fails to effectively reflect model performance when influenced by numerous false positives. To address this limitation, we further incorporate image-level and pixel-level AP (Area Under the Precision-Recall Curve) and F_1 max scores into our evaluation for a more comprehensive assessment.

Implementation details. During both training and testing phases, all images are resized to 512×512 pixels and center-cropped. Following common practices in previous literature, we select WideResNet50 [59] as the feature extractor. With network parameters frozen, we utilize the outputs from all blocks of layers 2 to 4 to compute global similarity features, and we select the outputs of the final block from each of layers 1 to 3 for nearest-neighbor feature searches. We employ the Adam [34] optimizer to update network parameters, setting the learning rate to 1×10^{-5} and weight decay to 1×10^{-4}. The total number of training epochs is set to 100, with a batch size of 4 and a random seed of 42. Similar to the training methodology of ResAD [55], we randomly select reference samples for each input image during training to enhance feature diversity. All experiments are conducted using a single NVIDIA RTX 4090 GPU.

Competing methods. Among traditional anomaly detection approaches, we select several classical full-shot methods and adapt them to few-shot settings, including SPADE [17], PaDiM [18], and PatchCore [42]. Additionally, we compare our approach with prompt-based methods, such as WinCLIP [32] and InCTRL [67]. Furthermore, we also include ResAD [55] and it shares a similar contrastive learning strategy with our method. Except for WinCLIP [32] and InCTRL [67] employing pretrained ViT-B-16 [21] as the backbone, all other methods utilize WideResNet50 [59] as the backbone with parameters frozen during the training phase. To ensure a fair comparison, we guarantee that all methods used the same normal samples during the testing phase.

5.2 Main Results

Challenges of our ReinAD dataset. Tab. 2 highlights the distinctive challenges of our ReinAD dataset compared to existing datasets. We first train all baselines on MVTec AD [6], then conduct 1-shot evaluation across VisA [68], BTAD [36], MPDD [33], and our ReinAD dataset. Notably, models evaluated on MVTec AD [6] are trained on VisA [68]. Experimental results reveal two key observations: (1) State-of-the-art methods have achieved strong performance on existing benchmarks: 93.7% Image-AUROC / 93.6% Pixel-AUROC on MVTec AD [6], 86.5% Image-AUROC / 95.5% Pixel-AUROC on VisA [68], and 92.3% Image-AUROC / 96.4% Pixel-AUROC on BTAD [36]. (2) However, the same methods suffer significantly reduced performance on the ReinAD dataset, with the best approach achieving only 69.0% Image-AUROC and 86.7% Pixel-AUROC. The substantial performance drop reveals that current datasets may oversimplify industrial scenarios. In contrast, our dataset contains unaligned samples and multi-class fine-grained anomalies, better matching real

Table 2: **Anomaly detection and localization results** under 1-shot setting. All models are trained on MVTecAD datasets then tested on multiple datasets. Metrics are AUROC / AP / F_1 max. The best and second-best results are **bold** and underlined, respectively.

	Datasets	Classical AD Methods			Prompt-based AD Methods		Compare-based Methods	
		SPADE [17]	**PaDiM** [18]	**PatchCore** [42] (CVPR2022)	**WinCLIP** [32] (CVPR2023)	**InCTRL** [67] (CVPR2024)	**ResAD** [55] (NIPS2024)	**ReinADNet** (Ours)
Image-level	MVTecAD [6]	72.2/86.6/87.5	74.5/86.5/88.7	82.6/91.9/91.7	**93.7/96.9/94.5**	88.5/93.8/91.5	84.3/92.7/90.7	85.6/93.1/89.8
	VisA [68]	73.0/77.5/78.4	53.2/60.4/73.8	74.4/78.4/78.9	79.9/81.8/81.3	75.9/78.7/78.1	80.3/83.8/80.4	**86.5/89.7/84.5**
	BTAD [36]	86.9/93.9/90.7	87.5/81.8/80.2	87.8/85.5/80.8	84.8/85.9/80.8	**92.3**/93.3/88.2	88.3/91.1/86.0	92.2/**97.8/94.8**
	MPDD [33]	57.4/66.3/75.8	50.0/61.3/74.9	56.4/63.0/77.0	**68.3/72.2/80.6**	66.0/71.7/78.8	65.6/68.1/79.7	67.7/71.7/78.0
	ReinAD (Ours)	59.7/48.6/58.9	55.9/48.8/58.3	60.3/52.4/61.4	**68.7**/59.5/62.6	59.2/53.0/61.0	64.9/55.0/62.5	68.0/**59.7/64.9**
Pixel-level	MVTecAD [6]	90.5/34.2/39.0	88.8/32.3/37.5	92.1/**44.2/48.0**	**93.6**/38.6/42.8	-	93.1/43.1/46.4	**93.6**/43.7/46.7
	VisA [68]	92.3/14.4/20.2	84.9/5.6/9.5	93.6/26.5/31.4	84.6/15.8/23.4	-	**95.5**/31.2/37.2	94.7/**33.2/39.0**
	BTAD [36]	95.6/33.5/42.5	94.4/29.9/37.5	94.0/30.0/37.0	95.6/43.6/49.6	-	95.5/41.8/46.2	**96.4/51.7/52.1**
	MPDD [33]	93.7/14.6/19.7	87.5/7.5/13.3	93.1/16.3/18.6	94.4/**30.3/31.8**	-	**95.3**/24.8/26.8	93.8/26.6/28.4
	ReinAD (Ours)	**86.7**/7.1/10.5	74.6/2.2/5.0	81.9/7.7/10.3	85.9/7.7/13.2	-	86.3/8.3/13.3	86.3/**10.7/15.4**

Table 3: **Anomaly detection and localization results**. All models are trained and then tested on our ReinAD dataset under 1/2/4-shot settings. Metrics are AUROC / AP / F_1 max. The best and second-best results are **bold** and underlined, respectively. Detailed results for each category are available in the supplementary material.

	Setting	ClassicalAD Methods			Prompt-based AD Methods		Compare-based Methods	
		SPADE [17]	**PaDiM** [18]	**PatchCore** [42] (CVPR2022)	**WinCLIP** [32] (CVPR2023)	**InCTRL** [67] (CVPR2024)	**ResAD** [55] (NIPS2024)	**ReinADNet** (Ours)
Image-level	1-shot	59.7/48.6/58.9	55.9/48.8/58.3	60.3/52.4/61.4	68.7/59.5/62.6	53.3/49.4/58.5	67.0/57.5/64.5	**71.2/64.5/67.6**
	2-shot	61.3/50.1/59.1	57.4/49.4/58.8	61.6/52.5/60.7	70.3/59.8/63.2	54.0/48.9/58.7	70.5/61.4/65.3	**72.0/65.1/68.0**
	4-shot	64.1/52.3/59.7	63.6/51.0/60.6	61.9/51.7/61.0	71.2/60.5/63.9	54.6/49.3/58.7	73.0/63.1/66.1	**73.8/66.2/68.0**
Pixel-level	1-shot	86.7/7.1/10.5	74.6/2.2/5.0	81.9/7.7/10.3	85.9/7.7/13.2	-	89.6/10.4/15.8	**90.2/15.6/20.4**
	2-shot	86.1/5.2/8.8	75.8/2.7/5.9	81.9/5.7/8.8	86.8/8.1/13.6	-	**91.0**/12.0/18.4	90.3/**16.3/20.8**
	4-shot	87.7/6.8/11	83.9/3.9/8.1	80.0/4.7/7.9	87.5/9.0/14.6	-	**91.9**/14.5/21.4	89.7/**16.6/22.3**

industrial scenes. Thus, the ReinAD dataset provides a more rigorous benchmark that encourages development of anomaly detection methods capable of handling real industrial challenges.

Generalization ablity of our ReinADNet method. Tab. 3 validates the generalization ability of our ReinADNet method through few-shot evaluation. All methods are trained on our ReinAD training set and evaluated on our ReinAD testing set under 1/2/4-shot settings. Experimental results reveal two critical insights: (1) Since our dataset requires normal-anomaly comparisons to identify anomalies, contrastive-based approaches (ResAD [55] and our method) dominate performance rankings across almost all settings and metrics. (2) Designed for unaligned samples and multi-class fine-grained anomalies, ReinADNet outperforms all baselines under almost all settings and metrics (*e.g.* 64.5% *v.s.* 59.5% in image-level AP under 1-shot setting). Our method achieves better generalization by treating anomaly detection as a contrastive learning paradigm rather than memorizing normal patterns. This contrastive ability can generalize to novel categories. Above results demonstrate our method's strengths of contrastive representation learning and cross-sample fine-grained alignment. Such capabilities are crucial for real-world industrial inspection scenarios.

Qualitative results. Fig. 5 shows qualitative results on our testing set under 1-shot setting. Most state-of-the-art methods fail to generate good anomaly localization maps for new classes, due to many false positives in normal regions. However, our method effectively reduces false positives in normal regions and locate anomalies more accurately. The LED, PCB solder and thread samples highlight our method's robust feature matching for unaligned regions, where traditional methods often fail. Additionally, the plastic cover case shows our method's exceptional sensitivity to fine-grained anomalies. It can detect subtle anomalies that baseline approaches typically miss. The visual results complement our quantitative results, confirming our ReinADNet's superiority in handling both misalignment and fine-grained anomalies.

5.3 Ablation Studies

Tab. 4 presents the individual and combined detection performance of each module in our method. Specifically, "Search" refers to using only the Patch Comparison Strategy, where the global features

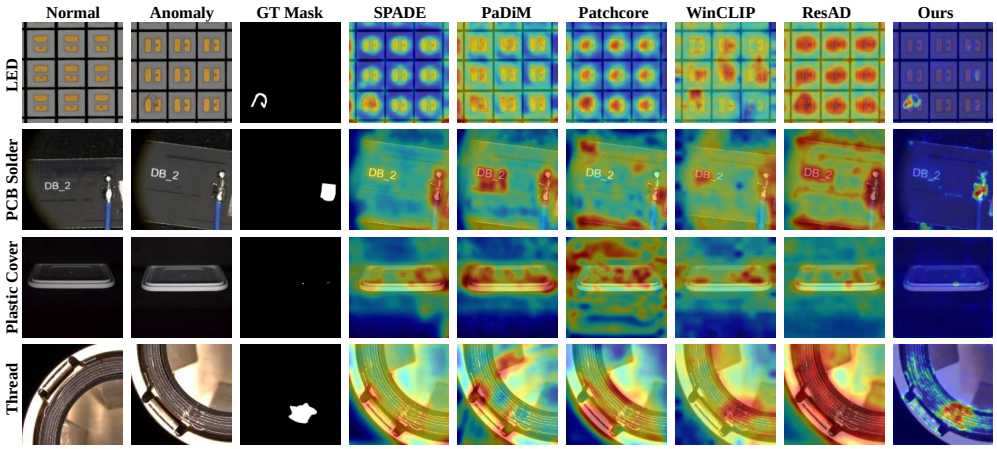


Figure 5: **Qualitative results**. More qualitative results are available in the Supplementary Material.

Table 4: **Ablation studies** of different network architectures. Metrics are AUROC / AP / F_1 max.

#	Method	Image-Level	Pixel-Level
0	Search	70.0 / 61.2 / 65.7	88.4 / 11.9 / 17.6
1	Aggregation	57.9 / 50.9 / 59.3	86.7 / 5.5 / 9.6
2	Aggregation+Query	66.8 / 60.1 / 63.8	88.1 / 11.8 / 16.9
3	Aggregation+Search	**71.2 / 64.5 / 67.6**	**90.2 / 15.6 / 20.4**

Table 5: **Ablation studies** of image-level anomaly score selection strategies.

	I-AUROC	I-AP	I-F_1 max
Maximum	**71.2**	**64.5**	**67.6**
Top 5%	70.2	62.5	65.4
Top 10%	69.2	60.5	64.0
Top 20%	68.0	58.5	62.5

output by the similarity aggregation module are removed, while the remaining structure is kept consistent with the full model. “Aggregation" denotes the use of only the Cost Aggregation Module, where the subsequent residual feature pyramid fusion module is excluded, and the global similarity features are directly decoded to produce the output. “Aggregation+Query" represents a variant of the model where the features involved in the Predictor module aggregation are the original query image features rather than residual features, with the rest of the structure identical to the full model. “Aggregation+Search" denotes the complete model configuration.

Analysis of module functionality. As illustrated in Tab. 4, local discrepancy features derived from nearest-neighbor search effectively complement global features obtained through similarity aggregation, thereby enhancing detection performance (*i.e.* #1 *v.s.* #3), additionally, the residual features derived by feature subtraction after search mitigate the category gap and demonstrate stronger generalization capabilities than the original query features (*i.e.* #2 *v.s.* #3). The cost aggregation module consolidates global contextual information across image pairs, further refining the local differential features (*i.e.* #0 *v.s.* #3).

Calculation of image-level anomaly scores. The image-level anomaly scores are directly derived from the output pixel-level anomaly score maps, rather than being produced by a separately trained network. Here we compare the image-level performance on our dataset using different strategies: the maximum value of the anomaly score map, and the average of the top n% highest scores in the entire map (with n set to 5, 10, and 20). As shown in Tab. 5, the best detection performance is achieved when using the maximum anomaly value as the image-level anomaly score, and as the value of n increases, the detection performance gradually declines. This indicates that the model effectively distinguishes between normal and anomalous instances, with a significant gap between the highest anomaly score and the scores of normal regions. In contrast, introducing the top n% averaging strategy dilutes the anomaly severity and led to reduced performance.

Contrastive-based *v.s.* zero-shot methods. To validate the advantages of contrastive-based methods, we compare them against several zero-shot methods [11, 14, 32] on our dataset. Our contrastive setting requires simultaneous input of both normal references and query images, while most existing zero-shot methods can only accept query images as inputs. Therefore, we only input query images to evaluate zero-shot methods. The results are given in Tab. 6, the zero-shot approaches demonstrate worse performance compared to contrastive-based methods. For instance, even for our ReinAD under 1-shot setting, the advantage over WinCLIP [32] is over 5% at image-level AUROC and 12%

Table 6: **Comparison** between zero-shot methods and contrastive-based methods under 1/2/4-shot settings. Metrics are AUROC / AP / F_1 max.

Shot	Method	Image-Level	Pixel-Level
0	APRIL-GAN [14]	61.8/55.9/60.3	78.7/6.4/11.7
	WinCLIP [32]	65.5/57.0/61.0	77.9/2.3/5.6
	AdaCLIP [11]	64.7/58.0/61.7	82.1/9.1/13.7
1	ResAD [55]	67.0/57.5/64.5	89.6/10.4/15.8
	ReinADNet (Ours)	71.2/64.5/67.6	90.2/15.6/20.4
2	ResAD [55]	70.5/61.4/65.3	91.0/12.0/18.4
	ReinADNet (Ours)	72.0/65.1/68.0	90.3/16.3/20.8
4	ResAD [55]	73.0/63.1/66.1	91.9/14.5/21.4
	ReinADNet (Ours)	73.8/66.2/68.0	89.7/16.6/22.3

Table 7: **Quantitative results** of supervised defect classification methods and unsupervised anomaly detection methods on our ReinAD dataset. Here we adopt AUROC / AP / F_1 max as evaluation metrics.

	Method	Image-Level	Pixel-Level
Sup.	DevNet [38]	69.0/85.1/86.0	-
	DRA [22]	75.6/91.2/89.6	-
Unsup.	SPADE [17]	75.4/88.7/88.2	85.5/7.4/12.0
	PaDiM [18]	81.9/91.5/91.0	92.4/18.3/25.0
	PatchCore [42]	83.7/92.5/91.0	92.6/19.2/24.7
	UniAD [57]	74.5/88.7/88.6	89.5/12.6/18.8

at pixel-level AUROC. As the number of shots increases, contrastive-based methods demonstrate greater advantages over zero-shot approaches.

5.4 Extended Applications of ReinAD

Beyond generalizable anomaly detection, our ReinAD dataset can be applied to extensive industrial anomaly detection tasks. First, it captures unaligned samples and multi-class fine-grained anomalies, better matching real-world complexity. Thus, it can be directly used for both one-for-one and one-for-many unsupervised anomaly detection methods. Second, as the largest real industrial dataset with pixel-level annotations, our ReinAD dataset enables backbone pre-training. Notably, the wide-used WideResNet50 [59] is pre-trained on ImageNet [20], exhibiting a critical domain gap with industrial scenarios. Therefore, a backbone pre-trained on a real industrial dataset can extract specific feature of industrial scenarios, improving the accuracy and generalization capability of IAD methods.

To demonstrating the broad applicability of our dataset, we evaluate two supervised [22, 38] and four unsupervised [17, 18, 42, 57] methods on our dataset. In these two settings, each category is split into training and test sets at an 8:1 ratio. We train the supervised models by classifying normal and anomaly samples, and then test on the same categories. The unsupervised AD methods are trained with only normal samples, and tested on the same categories. We conduct parts of experiments with the Ader [62] framework. Note that these experimental results cannot be compared with the results of few-shot methods before, since all the few-shot models are directly tested on categories unseen during training. Results in Tab. 7 demonstrate the usability of our dataset in both supervised and unsupervised settings.

6 Conclusion

We propose **ReinAD**, a comprehensive dataset for **Re**al-world **in**dustrial **A**nomaly **D**etection. Our dataset focuses on contrastive capability, containing unaligned samples and multi-class fine-grained anomalies. These features better match actual industrial scenarios. Based on our dataset, we introduce the ReinADNet method. Our method detects anomalies by comparing normal and anomaly samples, and can effectively identify fine-grained unaligned anomalies. Extensive experiments on ReinAD and several popular datasets demonstrate our dataset's challenge and our method's generalization ability.

Limitation and future work. While our ReinAD dataset offers the most diverse categories within existing industrial anomaly detection datasets, it still represents only a fraction of real industrial scenarios. Future work could extend coverage to more industrial categories, especially those with complex logical anomalies. Additionally, our method incurs higher computational costs due to its multi-scale matching approach. Thus, optimizing inference efficiency without sacrificing accuracy presents a key challenge.

7 Acknowledgements

This work was jointly supported by the National Natural Science Foundation of China (No.62422311), Shanghai Committee of Science and Technology, China (No.24TS1413500), the Fundamental Research Funds for the Central Universities (No.YG2025ZD07), and Shanghai Jiao Tong University 2030 Initiative.

References

[1] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon. Ganomaly: Semi-supervised anomaly detection via adversarial training. In *ACCV*, 2019.

[2] A. Arodi, M. Luck, J.-L. Bedwani, A. Zaimi, G. Li, N. Pouliot, J. Beaudry, and G. M. Caron. CableInspect-AD: An Expert-Annotated Anomaly Detection Dataset. In *NeurIPS*, 2024.

[3] A. Baitieva, D. Hurych, V. Besnier, and O. Bernard. Supervised Anomaly Detection for Complex Industrial Images. In *CVPR*, 2024.

[4] A. Baitieva, Y. Bouaouni, A. Briot, D. Ameln, S. Khalfaoui, and S. Akcay. Beyond academic benchmarks: Critical analysis and best practices for visual industrial anomaly detection. *arXiv preprint arXiv:2503.23451*, 2025.

[5] T. Bao, J. Chen, W. Li, X. Wang, J. Fei, L. Wu, R. Zhao, and Y. Zheng. MIAD: A maintenance inspection dataset for unsupervised anomaly detection. In *ICCV*, 2023.

[6] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. MVTec AD–A comprehensive real-world dataset for unsupervised anomaly detection. In *CVPR*, 2019.

[7] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger. Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings. In *CVPR*, 2020.

[8] J. Božič, D. Tabernik, and D. Skočaj. Mixed supervision for surface-defect detection: From weakly to fully supervised learning. *Computers in Industry*, 2021.

[9] T. Cao, J. Zhu, and G. Pang. Anomaly detection under distribution shift. In *ICCV*, 2023.

[10] Y. Cao, X. Xu, C. Sun, Y. Cheng, Z. Du, L. Gao, and W. Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023.

[11] Y. Cao, J. Zhang, L. Frittoli, Y. Cheng, W. Shen, and G. Boracchi. AdaCLIP: Adapting CLIP with Hybrid Learnable Prompts for Zero-Shot Anomaly Detection. In *ECCV*, 2024.

[12] L. Chen, Z. You, N. Zhang, J. Xi, and X. Le. UTRAD: Anomaly detection and localization with U-transformer. *Neural Networks*, 2022.

[13] L. Chen, Z. You, N. Zhang, J. Xi, and X. Le. Utrad: Anomaly detection and localization with u-transformer. *Neural Networks*, 2022.

[14] X. Chen, Y. Han, and J. 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.

[15] X. Chen, J. Zhang, G. Tian, H. He, W. Zhang, Y. Wang, C. Wang, and Y. Liu. CLIP-AD: A Language-Guided Staged Dual-Path Model for Zero-shot Anomaly Detection. *arXiv preprint arXiv:2311.00453*, 2024.

[16] S. Cho, S. Hong, S. Jeon, Y. Lee, K. Sohn, and S. Kim. Cats: Cost aggregation transformers for visual correspondence. *Advances in Neural Information Processing Systems*, 34:9011–9023, 2021.

[17] N. Cohen and Y. Hoshen. Sub-image anomaly detection with deep pyramid correspondences. *arXiv preprint arXiv:2005.02357*, 2020.

[18] T. Defard, A. Setkov, A. Loesch, and R. Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *ICLR*, 2021.

[19] H. Deng and X. Li. Anomaly detection via reverse distillation from one-class embedding. In *CVPR*, 2022.

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[22] M. Geppert, V. Larsson, J. L. Schönberger, and M. Pollefeys. Privacy preserving partial localization. In *CVPR*, 2022.

[23] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *AAAI*, 2024.

[24] D. Gudovskiy, S. Ishizaka, and K. Kozuka. Cflow-ad: Real-time unsupervised anomaly detection with localization via conditional normalizing flows. In *WACV*, 2022.

[25] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017.

[26] L. Heckler-Kram, J.-H. Neudeck, U. Scheler, R. König, and C. Steger. The MVTec AD 2 Dataset: Advanced Scenarios for Unsupervised Anomaly Detection. *arXiv preprint arXiv:2503.21622*, 2025.

[27] S. Hong, S. Cho, J. Nam, S. Lin, and S. Kim. Cost aggregation with 4d convolutional swin transformer for few-shot segmentation. In *European Conference on Computer Vision*, pages 108–126. Springer, 2022.

[28] J. Hou, Y. Zhang, Q. Zhong, D. Xie, S. Pu, and H. Zhou. Divide-and-assemble: Learning block-wise memory for unsupervised anomaly detection. In *ICCV*, 2021.

[29] Z. Hu and Z. Zhang. Sowa: Adapting hierarchical frozen window self-attention to visual-language models for better anomaly detection. *arXiv preprint arXiv:2407.03634*, 2024.

[30] C. Huang, H. Guan, A. Jiang, Y. Zhang, M. Spratling, and Y.-F. Wang. Registration Based Few-Shot Anomaly Detection. In *ECCV*, 2022.

[31] Y. Huang, C. Qiu, and K. Yuan. Surface defect saliency of magnetic tile. *The Visual Computer*, 2020.

[32] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *CVPR*, 2023.

[33] S. Jezek, M. Jonak, R. Burget, P. Dvorak, and M. 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 (ICUMT)*, 2021.

[34] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[35] X. Li, Z. Zhang, X. Tan, C. Chen, Y. Qu, Y. Xie, and L. Ma. Promptad: Learning prompts with only normal samples for few-shot anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16838–16848, 2024.

[36] P. Mishra, R. Verk, D. Fornasier, C. Piciarelli, and G. L. Foresti. VT-ADL: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, 2021.

[37] Y. Pan, L. Wang, Y. Chen, W. Zhu, B. Peng, and M. Chi. PA-CLIP: Enhancing zero-shot anomaly detection through pseudo-anomaly awareness. *arXiv preprint arXiv:2503.01292*, 2025.

[38] G. Pang, C. Ding, C. Shen, and A. v. d. Hengel. Explainable deep few-shot anomaly detection with deviation networks. *arXiv preprint arXiv:2108.00462*, 2021.

[39] H. Park, J. Noh, and B. Ham. Learning memory-guided normality for anomaly detection. In *CVPR*, 2020.

[40] Z. Qu, X. Tao, M. Prasad, F. Shen, Z. Zhang, X. Gong, and G. Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. *arXiv preprint arXiv:2407.12276*, 2024.

[41] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[42] K. Roth, L. Pemula, J. Zepeda, B. Schölkopf, T. Brox, and P. Gehler. Towards total recall in industrial anomaly detection. In *CVPR*, 2022.

[43] M. Salehi, N. Sadjadi, S. Baselizadeh, M. H. Rohban, and H. R. Rabiee. Multiresolution knowledge distillation for anomaly detection. In *CVPR*, 2021.

[44] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 2020.

[45] T. D. Tien, A. T. Nguyen, N. H. Tran, T. D. Huy, S. Duong, C. D. T. Nguyen, and S. Q. Truong. Revisiting reverse distillation for anomaly detection. In *CVPR*, 2023.

[46] C. Wang, W. Zhu, B.-B. Gao, Z. Gan, J. Zhang, Z. Gu, S. Qian, M. Chen, and L. Ma. Real-iad: A real-world multi-view dataset for benchmarking versatile industrial anomaly detection. In *CVPR*, 2024.

[47] G. Wang, S. Han, E. Ding, and D. Huang. Student-teacher feature pyramid matching for anomaly detection. *arXiv preprint arXiv:2103.04257*, 2021.

[48] S. Wang, L. Wu, L. Cui, and Y. Shen. Glancing at the patch: Anomaly localization with global and local feature comparison. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 254–263, 2021.

[49] X. Wang, J. Li, W. Zheng, H. Luo, F. Wang, Y. Chen, and R. Huang. A sim-to-real instance segmentation framework for densely stacked cartons. *IEEE Robotics and Automation Letters*, 2024.

[50] X. Wang, J. Li, W. Zheng, H. Luo, F. Wang, Y. Chen, and R. Huang. A sim-to-real instance segmentation framework for densely stacked cartons. *IEEE Robotics and Automation Letters*, 2024.

[51] T. Xiang, Y. Zhang, Y. Lu, A. L. Yuille, C. Zhang, W. Cai, and Z. Zhou. Squid: Deep feature in-painting for unsupervised anomaly detection. In *CVPR*, 2023.

[52] E. Yang, P. Xing, H. Sun, W. Guo, Y. Ma, Z. Li, and D. Zeng. 3CAD: A Large-Scale Real-World 3C Product Dataset for Unsupervised Anomaly. In *AAAI*, 2025.

[53] X. Yao, R. Li, Z. Qian, Y. Luo, and C. Zhang. Focus the discrepancy: Intra-and inter-correlation learning for image anomaly detection. In *ICCV*, 2023.

[54] X. Yao, C. Zhang, R. Li, J. Sun, and Z. Liu. One-for-all: Proposal masked cross-class anomaly detection. In *AAAI*, 2023.

[55] X. Yao, Z. Chen, C. Gao, G. Zhai, and C. Zhang. ResAD: A Simple Framework for Class Generalizable Anomaly Detection. In *NeurIPS*, 2024.

[56] X. Yao, R. Li, Z. Qian, L. Wang, and C. Zhang. Hierarchical gaussian mixture normalizing flow modeling for unified anomaly detection. In *ECCV*, 2024.

[57] Z. You, L. Cui, Y. Shen, K. Yang, X. Lu, Y. Zheng, and X. Le. A unified model for multi-class anomaly detection. In *NeurIPS*, 2022.

[58] Z. You, K. Yang, W. Luo, L. Cui, Y. Zheng, and X. Le. Adtr: Anomaly detection transformer with feature reconstruction. In *International Conference on Neural Information Processing*, 2022.

[59] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[60] M. Z. Zaheer, J.-h. Lee, M. Astrid, and S.-I. Lee. Old is gold: Redefining the adversarially learned one-class classifier training paradigm. In *CVPR*, 2020.

[61] V. Zavrtanik, M. Kristan, and D. Skočaj. Reconstruction by inpainting for visual anomaly detection. *PR*, 2021.

[62] J. Zhang, H. He, Z. Gan, Q. He, Y. Cai, Z. Xue, Y. Wang, C. Wang, L. Xie, and Y. Liu. A comprehensive library for benchmarking multi-class visual anomaly detection. *arXiv preprint arXiv:2406.03262*, 2024.

[63] T. Zhang, S. Zhong, W. Xu, L. Yan, and X. Zou. Catenary Insulator Defects Detection: A Dataset and an Unsupervised Baseline. *IEEE Transactions on Instrumentation and Measurement*, 2024.

[64] Z. Zhang, Z. Zhao, X. Zhang, C. Sun, and X. Chen. Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction. *Computers in Industry*, 2023.

[65] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022.

[66] Q. Zhou, G. Pang, Y. Tian, S. He, and J. Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023.

[67] J. Zhu and G. Pang. Toward generalist anomaly detection via in-context residual learning with few-shot sample prompts. In *CVPR*, 2024.

[68] Y. Zou, J. Jeong, L. Pemula, D. Zhang, and O. Dabeer. SPot-the-Difference Self-supervised Pre-training for Anomaly Detection and Segmentation. In *ECCV*, 2022.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Sec. 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experimental details are provided in Sec. 4 and Sec. 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Our dataset and codes will be made publicly available upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Details are provided in Sec. 4 and Sec. 5.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report the experimental results following the convention in anomaly detection research, the same as previous works.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The information on the computer resources is shown in Sec. 5.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We strictly adhere to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our work does not have direct negative societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We collect our data from actual industrial scenarios, thus does not have such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all original papers and make sure that our usage is legal.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Our dataset and code are well documented and the documentation is provided alongside the assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing experiments or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: LLM is used only for writing in our work.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.