RareCLIP: Rarity-aware Online Zero-shot Industrial Anomaly Detection

Jianfang He[1,2] Min Cao[3,*] Silong Peng[1] Qiong Xie[1,*]
[1]Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences
[3]Soochow University, Suzhou, China
{hejianfang2022, silong.peng, qiong.xie}@ia.ac.cn, mcao@suda.edu.cn

Abstract

Large vision-language models such as CLIP have made significant strides in zero-shot anomaly detection through prompt engineering. However, most existing methods typically process each test image individually, ignoring the practical rarity of abnormal patches in real-world scenarios. Although some batch-based approaches exploit the rarity by processing multiple samples concurrently, they generally introduce unacceptable latency for real-time applications. To mitigate these limitations, we propose RareCLIP, a novel online zero-shot anomaly detection framework that enables sequential image processing in real-time without requiring prior knowledge of the target domain. RareCLIP capitalizes on the zero-shot capabilities of CLIP and integrates a dynamic test-time rarity estimation mechanism. A key innovation of our framework is the introduction of a prototype patch feature memory bank, which aggregates representative features from historical observations and continuously updates their corresponding rarity measures. For each incoming image patch, RareCLIP computes a rarity score by aggregating the rarity measures of its nearest neighbors within the memory bank. Moreover, we introduce a prototype sampling strategy based on dissimilarity to enhance computational efficiency, as well as a similarity calibration strategy to enhance the robustness of rarity estimation. Extensive experiments demonstrate that RareCLIP attains state-of-the-art performance with 98.2% image-level AUROC on MVTec AD and 94.4% on VisA, while achieving a latency of 59.4 ms. Code is available at https://github.com/hjf02/RareCLIP.

1. Introduction

Industrial anomaly detection (AD) plays a critical role in intelligent manufacturing by enabling robust quality control and efficient production. Traditional unsupervised AD methods (also known as full-shot methods) [7, 27, 31, 35, 38, 40] have achieved strong performance by training on large datasets of normal images. However, recent research has shifted toward reducing data dependency. Few-shot AD approaches [13, 14, 23, 33, 37] tackle this challenge by requiring only a limited number of normal samples. Extending this trend, zero-shot AD methods [3, 6, 16, 20, 41] eliminate the need for any prior domain-specific training, enabling direct deployment on previously unseen products.

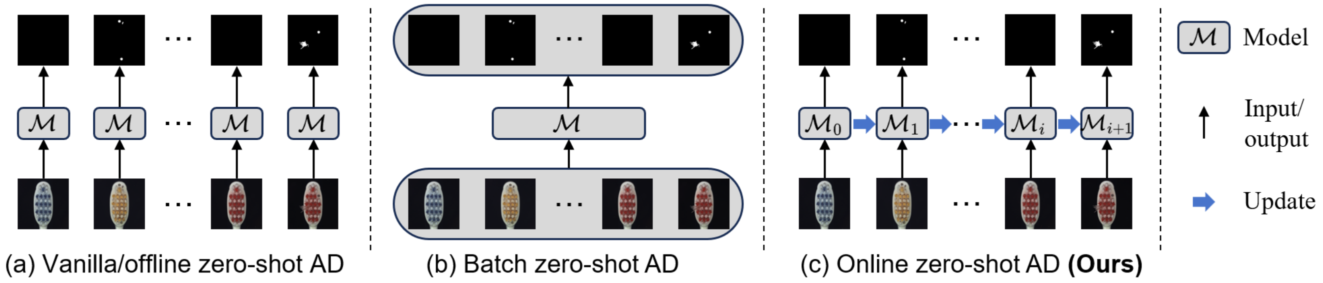
Most zero-shot AD methods use large vision-language models [17, 26, 30] and compare test images against predefined or learned textual prompts. However, by processing each test image in isolation, they overlook inter-image relationships. We refer to these as vanilla or offline zero-shot AD (Figure. 1a). They suffer from two major limitations: (1) They do not exploit the inherent rarity of anomalies in industrial environments, where defects are sparse compared to normal patterns; (2) Their isolated processing neglects the temporal continuity inherent in real-world production lines, where images are captured sequentially.

To overcome these limitations, recent work has explored batch zero-shot AD [18, 22], which models the distribution of an entire test dataset (Figure. 1b). However, these methods often rely on computationally expensive pairwise comparisons across all images to achieve high performance, resulting in prohibitive latency for time-sensitive applications.

In this paper, we introduce a novel paradigm: *online zero-shot AD*, where test images are processed sequentially in real-time (Figure. 1c). This setting introduces three key challenges: (1) **Online Modeling:** Rapid adaptation to data streams without prior domain knowledge. (2) **Single-Pass Efficiency:** Maximizing information extraction from images that are observed only once. (3) **Real-Time Performance:** Achieving low-latency and memory-efficient inference without compromising accuracy.

To address these challenges, we propose **RareCLIP**, the first online zero-shot AD framework that integrates the zero-shot capabilities of CLIP [30] with dynamic rarity modeling. Our key insight is that normal patches main-

[*] Corresponding authors.



(a) Vanilla/offline zero-shot AD (b) Batch zero-shot AD (c) Online zero-shot AD **(Ours)**

Figure 1. Illustration of zero-shot AD paradigms during testing. (a) Vanilla/Offline zero-shot AD processes each test image individually. (b) Batch zero-shot AD requires simultaneous input of test images. (c) Our proposed online zero-shot AD processes test images sequentially.

tain consistent appearances over time, whereas abnormal patches are rare and variable. We formalize this observation through the notion of *patch-image similarity*—defined as the maximum similarity between a given patch and all patches within an image. A patch is deemed normal if it exhibits high similarity to a sufficient fraction of historical images; otherwise, it is flagged as abnormal.

A direct implementation, RareCLIP-d, computes similarities against all historical images, which incurs substantial memory and computational overhead. To alleviate this, our proposed RareCLIP framework incorporates a prototype patch feature memory bank that continuously accumulates representative patch features while dynamically updating their rarity measures via a patch-image similarity memory bank. For each test patch, RareCLIP estimates its patch-image similarity by aggregating the patch-image similarities of its nearest neighbors within the patch feature memory bank. To efficiently identify representative patch features, we introduce a dissimilarity-based sampling strategy—*Sequential Coreset Sampling*—which effectively prunes redundant pairs from the memory bank. In addition, we propose a similarity calibration strategy, *Loose Similarity*, to mitigate estimation errors and enhance the robustness of rarity scoring.

Our main contributions are summarized as follows:

- We introduce online zero-shot AD, a novel and practical setting that addresses the limitations of offline and batch zero-shot paradigms in industrial anomaly detection.
- We propose RareCLIP, a pioneering framework that combines CLIP's zero-shot capability with dynamic rarity modeling. Key innovations include Sequential Coreset Sampling for efficient memory management and Loose Similarity for robust anomaly scoring.
- Extensive experiments on the MVTec AD and VisA datasets demonstrate that RareCLIP significantly outperforms existing methods.

2. Related Work

Industrial Anomaly Detection. Recent advances in anomaly detection have predominantly focused on unsupervised methods, also known as full-shot methods [1, 5, 7, 8, 11, 32, 38, 39], which rely on large collections of normal images during training. For example, PatchCore [31] builds a highly representative memory bank of normal patch features using a greedy coreset sampling strategy to minimize redundancy, thus reducing both storage requirements and inference time. In contrast, few-shot AD methods [13, 14, 23, 33, 37] address the challenge of limited normal data. GraphCore [37], for example, leverages graph neural networks to capture isometric invariant features, enabling the construction of a compact feature memory bank that performs effectively with only a few normal images.

Zero-shot Industrial Anomaly Detection. Leveraging the strong generalization capabilities of large pre-trained vision-language models [17, 26, 30], zero-shot AD methods have emerged that bypass the need for any normal reference images. Most zero-shot AD methods [4, 6, 12, 16, 41] use text prompts to compare against test images. WinCLIP [16] pioneered this direction by introducing a prompt ensemble strategy combined with multiple window-based forward passes across image patches. Subsequent works [4, 6, 12, 41] have primarily focused on improving the alignment between patch features and text prompts. We refer to these as vanilla or offline zero-shot AD methods since they typically process each test image individually. On the other hand, approaches such as ACR [18] and MuSc [22] explore batch zero-shot AD by jointly evaluating a collection of test images, although the high performance comes at the cost of increased latency that may hinder real-time application. In this paper, we propose OnlineAD, a framework for online zero-shot AD that leverages visual information from historical images in real-time.

Online Industrial Anomaly Detection. Online AD has attracted attention due to its closer alignment with industrial realities. Unlike incremental AD methods [25, 34], which iteratively refine models across multiple categories to enhance multi-class adaptability, online AD focuses on dynamically updating intra-class representations as new samples are acquired. For example, LeMO [10] supports online training using normal samples, making it well-suited

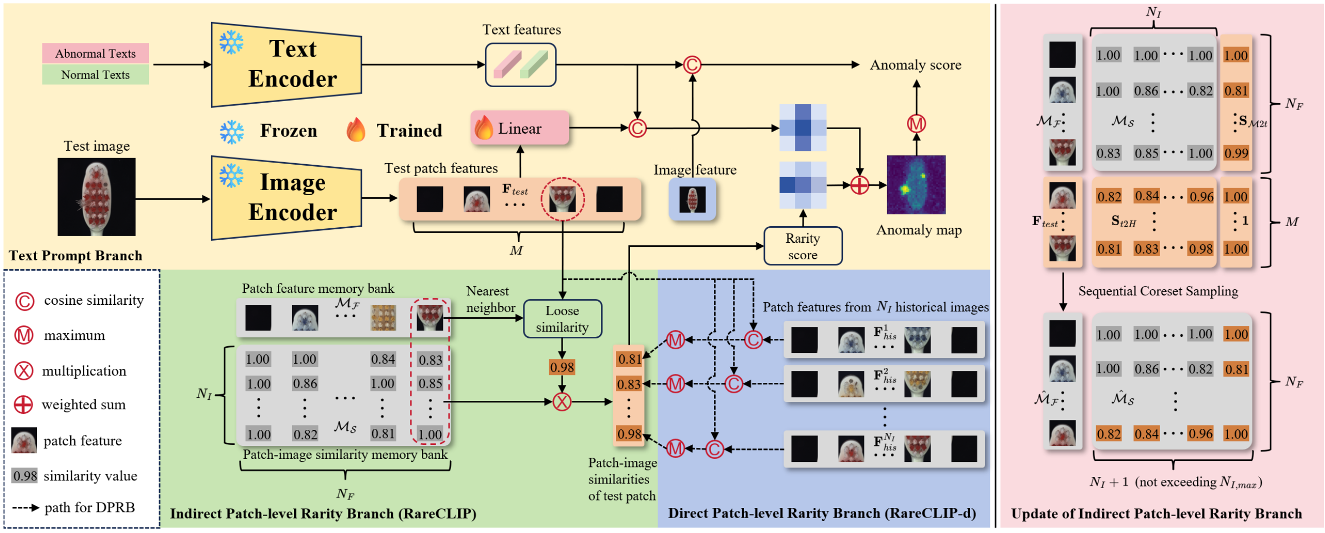


Figure 2. Overview of the proposed RareCLIP framework. RareCLIP mainly comprises two branches: the Text Prompt Branch (TPB, Section 3.1) and the Patch-level Rarity Branch. The latter includes the Direct Patch-level Rarity Branch (DPRB, Section 3.2.1) and the Indirect Patch-level Rarity Branch (IPRB, Section 3.2.2), where DPRB is an intermediate variant of IPRB. The right portion illustrates the IPRB update process, where $\mathbf{S}_{t2H}$ and $\mathbf{S}_{\mathcal{M}2t}$ denote the patch-image similarities between test patches and historical images, and between memorized patches and the test image, respectively.

for streaming industrial data. FOADS [36] advances online few-shot AD by constructing a normal feature bank from a limited number of normal samples and boost performance during online testing, while O-InReach [28] mitigates reliance on normal samples by delaying predictions until multiple samples have been observed. In contrast, our proposed online zero-shot AD framework can make predictions even in the absence of historical images.

3. Method

Problem Setting. Figure. 1(c) illustrates the formal setting for our proposed online zero-shot AD. Given an unlabeled stream of test images $D_u = \{I_1, I_2, \cdots\}$ from a single category without any prior domain-specific knowledge, the model is required to output the detection result for the t-th test image I_t before the next image I_{t+1} arrives.

Overview. Figure. 2 provides an overview of the proposed RareCLIP framework, which consists of two main branches: the *Text Prompt Branch* (TPB) and the *Patch-level Rarity Branch*. The Patch-level Rarity Branch is further divided into Direct Patch-level Rarity Branch (DPRB, corresponding to RareCLIP-d) and Indirect Patch-level Rarity Branch (IPRB, corresponding to RareCLIP), where DPRB serves as an ablation to highlight IPRB's reduced computational overhead while maintaining comparable performance. Each input test image is first encoded to generate patch features $\mathbf{F}_{test} = \{f^m_{test} \mid m \in [1, M]\} \in \mathbb{R}^{M \times C}$, where M is the number of patches and C is the feature dimension, and a global image feature F^{image}_{test}.

3.1. Text Prompt Branch

The Text Prompt Branch (TPB) forms the foundational component of RareCLIP, endowing it with initial zero-shot anomaly detection capabilities. In TPB, lightweight learnable projection layers adapt local patch features, denoted as $\widetilde{\mathbf{F}}_{test}$, to align effectively with text features $\mathbf{F}_{text}$ that encode normal and abnormal semantic information from fixed text prompt ensemble. The patch-level anomaly map is computed as:

$$\mathcal{A}_{text} = \text{softmax}(\langle \widetilde{\mathbf{F}}_{test}, \mathbf{F}_{text} \rangle), \tag{1}$$

where $\langle \cdot, \cdot \rangle$ denotes cosine similarity. Similarly, the global image feature F^{image}_{test} is compared with $\mathbf{F}_{text}$ to yield an image-level anomaly score:

$$c_{text} = \text{softmax}(\langle F^{image}_{test}, \mathbf{F}_{text} \rangle). \tag{2}$$

We observe that TPB tends to assign low anomaly scores in regions with missing components since background areas are generally learned as normal. To mitigate potential false negatives in such cases, we compute the cumulative average of $\mathcal{A}_{text}$ in the online setting, denoted as $\bar{\mathcal{A}}_{text}$, and then refine the result by taking the element-wise maximum:

$$\hat{\mathcal{A}}_{text} = \max(\mathcal{A}_{text}, \bar{\mathcal{A}}_{text}). \tag{3}$$

Training Loss of TPB. To align local patch features with the semantics of normal and abnormal text prompts, we employ lightweight learnable projection layers. During training, the adapted patch features $\widetilde{\mathbf{F}}$ are optimized using a combination of focal loss [24] and Dice loss [21] against the corresponding text features. To further mitigate overfitting, we also enforce alignment between $\widetilde{\mathbf{F}}$ and the global image feature—which is already well-aligned with text. The final loss is formulated as:

$$\begin{aligned}\text{Loss} = \ & \text{Focal}(\mathcal{A}_{text}, \mathcal{A}_{truth}) + \text{Dice}(\mathcal{A}_{text}, \mathcal{A}_{truth}) \\ & + \text{MSE}(\langle \widetilde{\mathbf{F}}, F^{image} \rangle, \mathbf{1}),\end{aligned} \tag{4}$$

where $\mathcal{A}_{truth}$ denotes the ground-truth anomaly mask and MSE represents the Mean Squared Error.

3.2. Patch-level Rarity Branch

The TPB provides semantic-aware anomaly priors, which are further refined by the Patch-level Rarity Branch through cross-sample pattern analysis.

3.2.1. Direct Patch-level Rarity Branch (RareCLIP-d)

To capture the rarity of each test patch, we introduce the Direct Patch-level Rarity Branch (DPRB), the direct implementation of rarity mechanism. In DPRB, patch features from N_I historical images are stored along with information about their source images to compute patch-image similarities. To manage memory usage, only patch features from the latest $N_{I,\text{max}}$ historical images are retained.

Patch-image Similarity Computation. For each test patch f_{test}^{m}, we compute a patch-image similarity vector:

$$\mathbf{s}_m = \left\{ \max_{f_{his} \in \mathbf{F}_{his}^{i}} \langle f_{test}^{m}, f_{his} \rangle \,\middle|\, i \in [1, N_I] \right\}, \tag{5}$$

where $\mathbf{F}_{his}^{i}$ denotes the patch features from the i-th historical image.

Rarity Score Computation. A patch is considered non-rare if a sufficient proportion (exceeding rarity threshold $X\%$) of its patch-image similarities yield high value. Given the patch-image similarity vector $\mathbf{s} \in \mathbb{R}^{N_I}$, we select the top $X\%$ of values, denoted as $\bar{\mathbf{s}}$, and define the rarity (anomaly) score as:

$$\text{Rarity}(\mathbf{s}) = 1 - \text{mean}(\bar{\mathbf{s}}). \tag{6}$$

Accordingly, the patch-level anomaly map is given by:

$$\mathcal{A}_{rare} = \left\{ \text{Rarity}(\mathbf{s}_m) \,\middle|\, m \in [1, M] \right\}. \tag{7}$$

Sequential Coreset Sampling. To reduce both memory and computational costs, we propose Sequential Coreset Sampling (SCS), which eliminates redundant features in the memory bank. The key idea is that if two features are highly similar, the latter can be removed with minimal loss of representativeness. This approach ensures that the remaining features are dissimilar from each other and cover the feature space with minimal redundancy.

SCS takes as input a feature set $\mathcal{F}$ and a target sample count N_{sample}. For each feature $\mathcal{F}_i$, we compute:

$$s_i = \begin{cases} 0, & i = 1, \\ \max\limits_{j<i} \langle \mathcal{F}_i, \mathcal{F}_j \rangle, & i > 1, \end{cases} \tag{8}$$

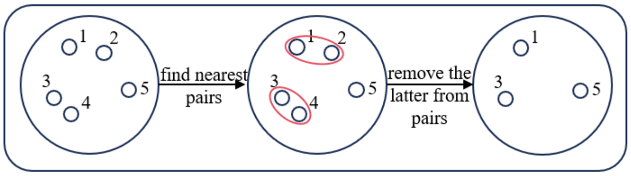


Figure 3. A simple example of Sequential Coreset Sampling. Numbers represent the sequence of features in the memory.

and then select the features with the smallest s_i values:

$$\mathcal{F}_{output} = \left\{ \mathcal{F}_j \,\middle|\, s_j \le s_{(N_{sample})} \right\} \in \mathbb{R}^{N_{sample} \times C}, \tag{9}$$

where $s_{(N_{sample})}$ is the N_{sample}-th smallest value. A simple example of SCS is shown in Figure. 3.

Within DPRB, we apply SCS to the patch features of each historical image, updating Eq. 5 as:

$$\mathbf{s}_m = \left\{ \max_{f_{his} \in \text{SCS}(\mathbf{F}_{his}^{i}, \alpha M)} \langle f_{test}^{m}, f_{his} \rangle \,\middle|\, i \in [1, N_I] \right\}, \tag{10}$$

where α is the sampling ratio.

3.2.2. Indirect Patch-level Rarity Branch (RareCLIP)

Although SCS reduces the number of features stored in DPRB, redundancy may still exist across different historical images. To address this, we propose the Indirect Patch-level Rarity Branch (IPRB) that employs dual memory banks:

- *Patch Feature Memory Bank* $\mathcal{M}_{\mathcal{F}} \in \mathbb{R}^{N_F \times C}$, which stores diversity-preserving patch features via SCS.
- *Patch-image Similarity Memory Bank* $\mathcal{M}_{\mathcal{S}} \in \mathbb{R}^{N_F \times N_I}$, which records the corresponding patch-image similarities of these memorized patch features.

The patch-image similarity for a test patch is then indirectly estimated by aggregating the similarities of its nearest neighbors in $\mathcal{M}_{\mathcal{F}}$ with their corresponding values in $\mathcal{M}_{\mathcal{S}}$.

Update of the Patch Feature Memory Bank. After processing a test image, its patch features $\mathbf{F}_{test}$ are appended to $\mathcal{M}_{\mathcal{F}}$. To maintain a fixed bank size, we apply SCS on the concatenated features:

$$\hat{\mathcal{M}_{\mathcal{F}}} = \text{SCS}\left(\text{Concat}(\mathcal{M}_{\mathcal{F}}, \mathbf{F}_{test}), N_F \right), \tag{11}$$

where $\hat{\mathcal{M}_{\mathcal{F}}}$ is the updated memory bank and N_F controls its size. Notably, SCS tends to preserve earlier features, ensuring minimal changes in $\mathcal{M}_{\mathcal{F}}$ and, consequently, reducing the frequency of estimated updates needed for the associated similarity bank $\mathcal{M}_{\mathcal{S}}$. This stability enhances the robustness of the patch-image similarity estimation.

Update of the Patch-image Similarity Memory Bank. The patch-image similarity memory bank $\mathcal{M}_{\mathcal{S}} \in \mathbb{R}^{N_F \times N_I}$ stores the similarity between each feature in $\mathcal{M}_{\mathcal{F}}$ and the N_I historical images. Its update can be divided into two components:

- *Memorized-to-Test Similarity* $\mathbf{S}_{\mathcal{M}2t} \in \mathbb{R}^{N_F}$: For each patch in $\mathcal{M}_{\mathcal{F}}$, we compute its patch-image similarity with the test image:

$$\mathbf{S}_{\mathcal{M}2t} = \left\{ \max_{f_{test} \in \mathbf{F}_{test}} \langle f^n_{\mathcal{M}}, f_{test} \rangle \,\middle|\, f^n_{\mathcal{M}} \in \mathcal{M}_{\mathcal{F}} \right\}. \quad (12)$$

- *Test-to-Historical Similarity* $\mathbf{S}_{t2H} \in \mathbb{R}^{M \times N_I}$: Since full historical images are unavailable in IPRB, we estimate $\mathbf{S}_{t2H}$ indirectly. For each test patch f^m_{test}, we first retrieve its K nearest neighbors $\{\phi^m_k\}^K_{k=1}$ from $\mathcal{M}_{\mathcal{F}}$ and compute their similarities:

$$\mathbf{S}_{t2\mathcal{M}} = \left\{ \langle f^m_{test}, \phi^m_k \rangle \,\middle|\, m \in [1, M],\, k \in [1, K] \right\} \in \mathbb{R}^{M \times K}. \quad (13)$$

  Then, for each test patch, we estimate its patch-image similarities to historical images by aggregating:

$$\mathbf{S}_{t2H} = \left\{ \frac{1}{K} \sum_{k=1}^{K} \mathbf{S}^{m,k}_{t2\mathcal{M}} \cdot \mathcal{M}_{\mathcal{S}}(\phi^m_k) \,\middle|\, m \in [1, M] \right\}, \quad (14)$$

  where $\mathbf{S}^{m,k}_{t2\mathcal{M}}$ is the similarity between f^m_{test} and its k-th neighbor ϕ^m_k, and $\mathcal{M}_{\mathcal{S}}(\phi^m_k) \in \mathbb{R}^{N_I}$ is the corresponding patch-image similarity vector of ϕ^m_k.

Using $\mathbf{S}_{t2H}$, the patch-level anomaly map can be obtained:

$$\mathcal{A}_{rare} = \left\{ \mathrm{Rarity}(\mathbf{s}_m) \,\middle|\, \mathbf{s}_m \in \mathbf{S}_{t2H} \right\}. \quad (15)$$

After computing $\mathbf{S}_{\mathcal{M}2t}$ and $\mathbf{S}_{t2H}$, the similarity memory bank is preliminarily updated as:

$$\mathcal{M}^*_{\mathcal{S}} = \begin{bmatrix} \mathcal{M}_{\mathcal{S}} & \mathbf{S}_{\mathcal{M}2t} \\ \mathbf{S}_{t2H} & \mathbf{1} \end{bmatrix} \in \mathbb{R}^{(N_F+M)\times(N_I+1)}, \quad (16)$$

where $\mathbf{1}$ is a vector of ones. When SCS is applied to $\mathcal{M}_{\mathcal{F}}$, the corresponding rows in $\mathcal{M}^*_{\mathcal{S}}$ associated with removed features are also deleted, resulting in an updated bank $\hat{\mathcal{M}}_{\mathcal{S}} \in \mathbb{R}^{N_F \times (N_I+1)}$. As before, we only retain patch-image similarities for the latest $N_{I,\max}$ historical images.

Loose Similarity. Multiplying similarity measures to estimate patch-image similarity in Eq. 14 can introduce error. Assuming that the test patch and its neighbor are both similar to the same patch in a historical image, this can be simplified to estimating $\cos\theta_{AC}$ as $\cos\theta_{AB} \cdot \cos\theta_{BC}$, where θ_{AC} is the angle between vectors A and C. The error $\cos\theta_{AC} - \cos\theta_{AB} \cdot \cos\theta_{BC}$ is bounded by

$$\pm\sqrt{(1-\cos^2\theta_{AB})(1-\cos^2\theta_{BC})},$$

which decreases as either $\cos\theta_{AB}$ or $\cos\theta_{BC}$ approaches 1. However, subtle differences among different images often prevent normal patches from achieving a similarity of 1, thus gradually reducing their estimated similarity after multiple multiplicative estimations.

To alleviate this issue and enhance the robustness of rarity estimation, we propose a *Loose Similarity* (LS) strategy that loosens the requirement for similarities to reach 1. Given a set of similarities $\mathbf{s} = \{s_1, s_2, \cdots, s_n\}$, we identify the $Y\%$-th largest value $s_{[p_Y]}$, where $p_Y = \lceil n \times Y\% \rceil$ and Y controls the loose degree. We then divide all values by $s_{[p_Y]}$ and cap the results at 1:

$$\mathrm{LS}(\mathbf{s}) = \min\left(\frac{\mathbf{s}}{s_{[p_Y]}}, 1\right). \quad (17)$$

We apply LS to both $\mathbf{S}_{\mathcal{M}2t}$ (Eq. 12) and $\mathbf{S}_{t2\mathcal{M}}$ (Eq. 13), ensuring symmetric error compensation:

$$\hat{\mathbf{S}}_{\mathcal{M}2t} = \mathrm{LS}(\mathbf{S}_{\mathcal{M}2t}), \quad \hat{\mathbf{S}}_{t2\mathcal{M}} = \mathrm{LS}(\mathbf{S}_{t2\mathcal{M}}). \quad (18)$$

3.3. Anomaly Detection

Image-level Rarity Branch. To better detect large abnormal regions, we introduce an Image-level Rarity Branch (IRB). In IRB, a local-aggregated image level feature F^{laif} is calculated to assess image-level rarity. Since TPB tends to yield low anomaly scores in monotonous areas, we use $\bar{\mathcal{A}}_{text}$ to roughly identify the most noticeable regions. Specifically, F^{laif} is obtained by averaging the top half of the patch features corresponding to the highest values in $\bar{\mathcal{A}}_{text}$. The similarity between the test image and each historical image is then computed as:

$$\mathbf{s}_{t2H} = \left\{ \langle F^{laif}_{test}, F^{laif}_i \rangle \,\middle|\, i \in [1, N_I] \right\}, \quad (19)$$

where F^{laif}_{test} and F^{laif}_i are the local-aggregated features for the test image and the i-th historical image, respectively. The image-level rarity (anomaly) score is then computed as $c_{rare} = \mathrm{Rarity}(\mathbf{s}_{t2H})$, and combined with the patch-level anomaly map:

$$\hat{\mathcal{A}}_{rare} = \mathcal{A}_{rare} + c_{rare} \cdot \Big(1 + \mathrm{normalize01}(\bar{\mathcal{A}}_{text})\Big). \quad (20)$$

Pixel-level Anomaly Detection. We fuse the refined anomaly maps from TPB and the rarity branch to obtain the final anomaly map:

$$\mathcal{A}_{test} = \lambda_{text}\hat{\mathcal{A}}_{text} + \lambda_{rare}\hat{\mathcal{A}}_{rare}, \quad (21)$$

where λ_{text} and λ_{rare} are weighting factors. The anomaly map $\mathcal{A}_{test} \in \mathbb{R}^M$ is reshaped to a $\sqrt{M} \times \sqrt{M}$ grid and upsampled to the original image resolution. A Gaussian filter is then applied to smooth the final pixel-level anomaly detection result.

Image-level Anomaly Detection. For image-level detection, we first compute a preliminary anomaly score c_{test} by taking the maximum value from the pixel-level anomaly map and adding c_{text}. Inspired by MuSc [22], we further refine this score via an Image-level Re-scoring strategy. The underlying assumption is that visually similar images should exhibit consistent anomaly statuses. We identify the B most similar historical images based on the average similarity of their image-level features (including both F^{image} and F^{laif}). Let s_b denotes the similarity between the test image and the b-th similar historical image, and c_b its preliminary anomaly score, for $b = 1, \ldots, B$. The refined image-level anomaly score is computed as:

$$\hat{c}_{test} = \frac{e^{\frac{1}{\tau}} \cdot c_{test} + \sum_{b=1}^{B} e^{\frac{s_b}{\tau}} \cdot c_b}{e^{\frac{1}{\tau}} + \sum_{b=1}^{B} e^{\frac{s_b}{\tau}}}, \tag{22}$$

where τ is a temperature hyper-parameter.

4. Experiments

Dataset. We conduct experiments on the widely-used industrial image datasets, MVTec AD [2] and VisA [42]. Both datasets contain multiple subsets, each with only one category per subset. MVTec AD includes 10 object categories and 5 texture categories with high-resolution images (from 700×700 to 1024×1024). VisA consists of 12 object categories with high-resolution images (1000×1500). We use the official test splits of both datasets, which include normal and abnormal images.

Evaluation Metrics. In line with common practice, we report three metrics for image-level anomaly detection: image-level Area Under Receiver Operator Characteristic curve (I-AUC), image-level F1-score at the optimal threshold (I-F1-max), and image-level Average Precision (I-AP). Additionally, we report four metrics for pixel-level anomaly detection: pixel-level Area Under Receiver Operator Characteristic curve (P-AUC), pixel-level F1-score at the optimal threshold (P-F1-max), pixel-level Average Precision (P-AP), and Per-Region Overlap (PRO).

Implementation Details. In accordance with prior studies [6, 41], we adopt CLIP [30] with the ViT-L-14-336 [9] backbone implemented via OpenCLIP [15] and resize all input images to 518×518. For the text prompt branch, features are extracted from the 12th, 16th, 20th, and 24th layers of ViT which contain rich semantic information, and the associated linear layers are trained on one dataset when testing on another. In the rarity branch, features from the 6th, 12th, 18th, and 24th layers are used, with multi-scale patch features obtained via 1×1 and 3×3 neighborhood average pooling to enhance visual representation. The rarity threshold $X\%$ is set to 30%, and $N_{I,\max}$ is fixed at 200 by default. In the Direct Patch-level Rarity Branch (DPRB), the sampling ratio α is set to $\frac{1}{3}$. In the Indirect Patch-level Rarity Branch (IPRB), the patch feature memory bank size N_F is set to 4107, K is fixed at 3 for K-NN, and the loose degree $Y\%$ is set to 1%. All experiments are performed on a single NVIDIA GeForce RTX 3090. Further implementation details are provided in the supplemental material.

Offline and Online Modes. RareCLIP can operate in both offline and online modes. In offline mode, all memory banks are frozen and cumulative updates are halted, ensuring that detection results are independent of the test sequence. In online mode, the model dynamically updates its memory banks, and performance may be influenced by the sequence of test images. To ensure reproducibility of online mode, we use the official test split as fixed initial sequence, then report the average performance over five different random shuffles (using seeds 0, 1, 2, 3, and 4).

Baselines. We compare the proposed RareCLIP with other state-of-the-art zero-shot AD methods, including: (1) offline zero-shot AD methods WinCLIP [16], April-GAN [6], and AnomalyCLIP [41], FiLo [12], AdaCLIP [4] and VCP-CLIP [29]; (2) batch zero-shot AD methods ACR [18] and MuSc [22]; (3) pseudo online (online-) zero-shot AD methods O-InReach [28] and MuSc*, which only make prediction starting from the second image, and MuSc* is a reproduced online version of MuSc [22] that only utilizes historical information from previous test images.

4.1. Comparison with Zero-shot Methods

Table. 1 presents the performance comparison between RareCLIP and other zero-shot AD methods. Our key observations are: (1) In offline mode, RareCLIP achieves competitive results compared to existing offline zero-shot methods; (2) In online mode, RareCLIP exhibits significant improvements across all metrics, even outperforming batch zero-shot AD methods; (3) RareCLIP attains higher image-level performance than RareCLIP-d (which incurs higher memory and computation costs) while only incurring a slight decrease in pixel-level performance. These results underscore the effectiveness and practicality of the proposed RareCLIP.

4.2. Comparison with Few-/Full-shot Methods

Table. 2 compares RareCLIP against state-of-the-art few-shot and full-shot AD methods [6, 16, 19, 23, 31, 35, 36, 40]. We also extend RareCLIP to a few-shot AD setting (details are provided in the supplemental material). Compared to state-of-the-art few-shot methods, RareCLIP in offline mode achieves more than a 1% improvement in both I-AUC and P-AUC on both datasets. Moreover, RareCLIP in online mode attains competitive performance with full-shot AD methods.

4.3. Offline Performance vs. Online Adaptation

To analyze the evolution of the model's performance as it processes more images, we update the model using the first N images of the test set and then evaluate its performance on the entire test set in offline mode. This sim-

Dataset	Method	Public	Mode	Image-level			Pixel-level			
				I-AUC	I-F1-max	I-AP	P-AUC	P-F1-max	P-AP	PRO
MVTec AD	WinCLIP [16]	CVPR23	offline	91.8	92.9	96.5	85.1	31.7	-	64.6
	April-GAN [6]	CVPR23 WS	offline	86.1	90.5	93.5	87.6	43.3	40.8	44.0
	AnomalyCLIP [41]	ICLR24	offline	91.6	92.7	96.4	91.1	39.1	34.5	81.4
	FiLo [12]	ACM MM24	offline	91.2	-	-	92.3	-	-	-
	AdaCLIP [4]	ECCV24	offline	90.0	92.3	95.7	89.9	43.9	41.6	44.1
	VCP-CLIP [29]	ECCV24	offline	92.1	91.8	96.9	92.0	49.3	49.4	87.3
	RareCLIP	-	offline	91.5	92.9	96.6	91.5	47.5	46.1	86.2
	O-InReach [28]	ECCV24	online-	87.1±0.9	91.1±0.4	94.0±0.5	93.5±0.1	43.9±0.4	37.7±0.3	83.0±0.4
	MuSc* [22]	ICLR24	online-	96.0±0.4	96.2±0.3	98.0±0.3	97.0±0.2	57.8±0.5	55.9±0.9	93.3±0.1
	RareCLIP-d	-	online	97.9±0.2	97.4±0.2	99.1±0.2	**97.8±0.1**	62.9±0.5	64.2±1.1	93.4±0.1
	RareCLIP	-	online	**98.2±0.2**	**97.6±0.1**	**99.3±0.1**	97.7±0.2	**64.1±0.7**	**66.1±1.0**	**93.5±0.1**
	ACR [18]	NIPS23	batch	85.8	91.3	92.9	92.5	44.2	38.9	72.7
	MuSc [22]	ICLR24	batch	97.7	97.5	99.1	97.1	62.2	62.3	93.4
VisA	WinCLIP [16]	CVPR23	offline	78.1	79.0	81.2	79.6	14.8	-	56.8
	April-GAN [6]	CVPR23 WS	offline	78.0	78.7	81.4	94.2	32.3	25.7	86.8
	AnomalyCLIP [41]	ICLR24	offline	82.0	80.4	85.3	95.5	28.3	21.3	86.7
	FiLo [12]	ACM MM24	offline	83.9	-	-	95.9	-	-	-
	AdaCLIP [4]	ECCV24	offline	85.8	83.1	88.5	95.5	37.7	31.5	51.3
	VCP-CLIP [29]	ECCV24	offline	83.8	81.4	87.6	95.7	29.8	30.1	90.7
	RareCLIP	-	offline	86.1	83.1	89.0	95.7	33.5	27.0	90.2
	O-InReach [28]	ECCV24	online-	78.0±0.2	79.4±0.3	82.3±0.3	95.7±0.1	31.4±0.9	25.5±1.1	75.7±0.6
	MuSc* [22]	ICLR24	online-	90.0±0.5	87.1±0.7	90.5±0.3	98.6±0.0	48.5±0.2	44.9±0.3	92.4±0.1
	RareCLIP-d	-	online	93.5±0.2	90.2±0.3	94.1±0.5	**98.9±0.0**	48.1±0.4	44.2±0.5	93.2±0.1
	RareCLIP	-	online	**94.4±0.3**	**90.8±0.4**	**95.3±0.2**	98.8±0.0	**50.9±0.3**	**47.5±0.3**	**93.5±0.1**
	MuSc [22]	ICLR24	batch	92.6	89.1	93.3	98.7	48.9	45.4	92.4

Table 1. Comparison of image-level and pixel-level zero-shot anomaly detection on the MVTec AD and VisA datasets. We compare the proposed RareCLIP with other state-of-the-art zero-shot methods. MuSc* denotes a reproduced online version of MuSc [22], and the evaluations of methods in "online-" mode exclude the first image since they predict result starting from the second image. Bold indicates the best performance, while underline denotes the second-best result. Methods under the batch setting are highlighted in gray as they concurrently utilize all test images for anomaly detection. All metrics are in %.

Method	Setting	Mode	MVTec AD		VisA	
			I-AUC	P-AUC	I-AUC	P-AUC
WinCLIP+ [16]	4-shot	offline	95.2	96.2	87.3	97.2
April-GAN [6]	4-shot	offline	92.8	95.9	92.6	96.2
PromptAD [23]	4-shot	offline	96.6	96.5	89.1	97.4
FOADS [36]	10-shot	online	87.3	95.1	-	-
RareCLIP	0-shot	online	98.2	97.7	94.4	98.8
RareCLIP	4-shot	offline	97.7	98.1	94.6	98.8
RareCLIP	4-shot	online	98.7	98.2	95.5	**98.8**
CutPaste [19]	full-shot	offline	96.1	96.0	-	-
PatchCore [31]	full-shot	offline	99.1	98.1	94.8	98.5
RD++ [35]	full-shot	offline	99.4	98.3	95.9	98.7
RealNet [40]	full-shot	offline	**99.6**	**99.0**	**97.8**	**98.8**

Table 2. Comparison with state-of-the-art few-shot and full-shot methods in image-level and pixel-level AUC on the MVTec AD and VisA datasets.

ulates placing each test image at time step $N + 1$ and allows us to observe how performance evolves with the amount of historical data. As shown in Figure. 3, we observe that: (1) Performance steadily improves as N increases, stabilizing after approximately $N = 24$ images; (2) RareCLIP achieves slightly higher image-level performance but marginally lower pixel-level performance compared to RareCLIP-d; (3) RareCLIP outperforms MuSc* especially when N is small, demonstrating its stronger adaptability in online setting.

4.4. Ablation Study

Sampling Strategy. We compared SCS with other sampling strategies, including Random Sampling (RS), K-means Clustering Sampling (KCS), and Greedy Coreset Sampling (GCS). We use the implementation of GCS from Patchcore [31]. The results in Table. 3 indicate that: (1) KCS yields performance similar to RS but requires significantly more memory and time; (2) GCS achieves higher image-level performance than RS, though at the cost of increased computational time; (3) SCS offers optimal performance with minimal additional memory and time overhead by maintaining minimal changes in memory banks.

Impact of $N_{I,max}$ on Memory and Time Cost. Table. 4 illustrates the effect of varying $N_{I,\max}$ on GPU memory and computation time. We observe that: (1) Both MuSc* and RareCLIP-d exhibit a significant increase in time and GPU memory usage as $N_{I,\max}$ increases due to their reliance on

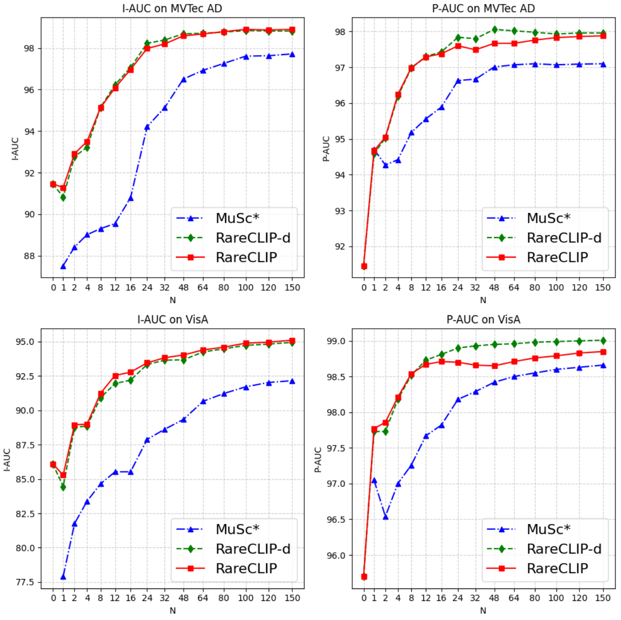


Figure 4. Offline performance as a function of the number of processed test images (N) on the MVTec AD and VisA datasets.

Sampling	MVTec AD		VisA		GPU (MB)	Time (ms)
	I-AUC	P-AUC	I-AUC	P-AUC		
RS	88.13	93.21	75.38	93.47	**4296**	**57.0**
KCS	88.05	93.10	75.38	93.35	4870	4280.8
GCS	97.18	93.17	87.80	88.46	4468	2197.6
SCS	**98.19**	**97.70**	**94.40**	**98.80**	4352	59.4

Table 3. Ablation study of different sampling strategies on the MVTec AD and VisA datasets. RS, KCS, GCS and SCS respectively denote Random Sampling, K-means Clustering Sampling, Greedy Coreset Sampling and Sequential Coreset Sampling.

$N_{I,max}$	MuSc* [22]		**RareCLIP-d**		**RareCLIP**	
	GPU (MB)	Time (ms)	GPU (MB)	Time (ms)	GPU (MB)	Time (ms)
50	4697	455.0	4696	71.6	4298	59.0
200	7293	983.7	6018	106.2	4352	59.4
1000	19857	4054.7	19586	310.8	4520	61.2

Table 4. Comparison of time and GPU memory consumption for MuSc* [22], and RareCLIP(-d) under different values of $N_{I,\text{max}}$.

storing extensive patch features from historical images; (2) RareCLIP maintains low time and memory requirements across varying values of $N_{I,\text{max}}$, highlighting its computational efficiency.

$Y(\%)$	MVTec AD		VisA	
	I-AUC	P-AUC	I-AUC	P-AUC
w/o	97.00	96.75	93.24	98.32
0	97.40	97.02	93.73	98.49
0.5	98.14	97.63	**94.45**	98.78
1	**98.19**	97.70	94.40	**98.80**
2	98.15	97.75	94.25	98.79
3	98.12	97.78	94.12	98.78
5	98.00	**97.79**	93.92	98.76

Table 5. Ablation study of loose degree Y on the MVTec AD and VisA datasets. "w/o" indicates that LS is not applied.

$X(\%)$	MVTec AD		VisA	
	I-AUC	P-AUC	I-AUC	P-AUC
5	97.24	96.68	93.57	98.82
10	97.60	96.88	94.08	**98.83**
20	98.15	97.31	94.36	98.81
30	**98.19**	97.70	**94.40**	98.80
40	97.98	**97.80**	94.12	98.76
50	97.63	97.78	93.90	98.74
60	97.26	97.76	93.62	98.71

Table 6. Ablation study of rarity threshold X on the MVTec AD and VisA datasets.

Impact of Loose Degree Y. Table. 5 presents an ablation study on the hyper-parameter Y, which controls the loose degree in LS. Our observations include: (1) Without LS or with $Y = 0$ (i.e., using the maximum similarity), performance is lower compared to using other Y values; (2) There is little difference in performance between $Y = 0.5$ and $Y = 2$; (3) As Y increases further, I-AUC performance tends to decrease due to more weak abnormal regions being misclassified as normal.

Impact of Rarity Threshold X. Table. 6 shows how varying the rarity threshold X affects performance. The results indicate that both excessively high and low values of X lead to degraded performance, while a moderate value yields the best results.

5. Conclusion

To tackle the novel online zero-shot AD task, we proposed RareCLIP, a rarity-aware method that leverages the zero-shot capabilities of CLIP and integrates a dynamic test-time rarity estimation mechanism. RareCLIP contains a prototype patch feature memory bank, which aggregates representative features from historical observations and continuously updates their rarity measures. For each test patch, RareCLIP computes a rarity score by aggregating the rarity measures of its nearest neighbors within the memory bank. Furthermore, we introduced a dissimilarity-based prototype sampling strategy—*Sequential Coreset Sampling*—to improve computational efficiency, along with a similarity calibration mechanism, *Loose Similarity*, to enhance the robustness of rarity estimation. Experimental results on benchmark datasets demonstrate that RareCLIP achieves state-of-the-art performance with low latency.

Acknowledgments

This work is supported in part by the National Natural Science Foundation of China under Grant No 62476188; in part by the Natural Science Foundation of the JiangsuHigher Education Institutions of China.

References

[1] Jaehyeok Bae, Jae-Han Lee, and Seyun Kim. Pni: Industrial anomaly detection using position and neighborhood information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6373–6383, 2023. 2

[2] 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*, pages 9592–9600, 2019. 6

[3] Yunkang Cao, Xiaohao Xu, Chen Sun, Yuqi Cheng, Zongwei Du, Liang Gao, and Weiming Shen. Segment any anomaly without training via hybrid prompt regularization. *arXiv preprint arXiv:2305.10724*, 2023. 1

[4] 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, 6, 7

[5] Qiyu Chen, Huiyuan Luo, Chengkan Lv, and Zhengtao Zhang. A unified anomaly synthesis strategy with gradient ascent for industrial anomaly detection and localization. *arXiv preprint arXiv:2407.09359*, 2024. 2

[6] Xuhai Chen, Yue Han, and Jiangning Zhang. 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. 1, 2, 6, 7

[7] Thomas Defard, Aleksandr Setkov, Angelique Loesch, and Romaric Audigier. Padim: a patch distribution modeling framework for anomaly detection and localization. In *International Conference on Pattern Recognition*, pages 475–489. Springer, 2021. 1, 2

[8] Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9737–9746, 2022. 2

[9] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[10] Han Gao, Huiyuan Luo, Fei Shen, and Zhengtao Zhang. Towards total online unsupervised anomaly detection and localization in industrial vision. *arXiv preprint arXiv:2305.15652*, 2023. 2

[11] Zhihao Gu, Liang Liu, Xu Chen, Ran Yi, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Annan Shu, Guannan Jiang, and Lizhuang Ma. Remembering normality: Memory-guided knowledge distillation for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16401–16409, 2023. 2

[12] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Hao Li, Ming Tang, and Jinqiao Wang. Filo: Zero-shot anomaly detection by fine-grained description and high-quality localization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2041–2049, 2024. 2, 6, 7

[13] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1932–1940, 2024. 1, 2

[14] Chaoqin Huang, Haoyan Guan, Aofan Jiang, Ya Zhang, Michael Spratling, and Yan-Feng Wang. Registration based few-shot anomaly detection. In *European Conference on Computer Vision*, pages 303–319. Springer, 2022. 1, 2

[15] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 6

[16] 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, 2, 6, 7

[17] 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. 1, 2

[18] Aodong Li, Chen Qiu, Marius Kloft, Padhraic Smyth, Maja Rudolph, and Stephan Mandt. Zero-shot anomaly detection via batch normalization. *Advances in Neural Information Processing Systems*, 36, 2023. 1, 2, 6, 7

[19] Chun-Liang Li, Kihyuk Sohn, Jinsung Yoon, and Tomas Pfister. Cutpaste: Self-supervised learning for anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9664–9674, 2021. 6, 7

[20] Shengze Li, Jianjian Cao, Peng Ye, Yuhan Ding, Chongjun Tu, and Tao Chen. Clipsam: Clip and sam collaboration for zero-shot anomaly segmentation. *arXiv preprint arXiv:2401.12665*, 2024. 1

[21] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019. 3

[22] Xurui Li, Ziming Huang, Feng Xue, and Yu Zhou. Musc: Zero-shot industrial anomaly classification and segmentation with mutual scoring of the unlabeled images. In *International Conference on Learning Representations*, 2024. 1, 2, 6, 7, 8

[23] Xiaofan Li, Zhizhong Zhang, Xin Tan, Chengwei Chen, Yanyun Qu, Yuan Xie, and Lizhuang 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. 1, 2, 6, 7

[24] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[25] Jiaqi Liu, Kai Wu, Qiang Nie, Ying Chen, Bin-Bin Gao, Yong Liu, Jinbao Wang, Chengjie Wang, and Feng Zheng. Unsupervised continual anomaly detection with contrastively-learned prompt. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3639–3647, 2024. 2

[26] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2

[27] 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. 1

[28] Declan McIntosh and Alexandra Branzan Albu. Unsupervised, online and on-the-fly anomaly detection for non-stationary image distributions. In *European Conference on Computer Vision*, pages 428–445. Springer, 2024. 3, 6, 7

[29] Zhen Qu, Xian Tao, Mukesh Prasad, Fei Shen, Zhengtao Zhang, Xinyi Gong, and Guiguang Ding. Vcp-clip: A visual context prompting model for zero-shot anomaly segmentation. *arXiv preprint arXiv:2407.12276*, 2024. 6, 7

[30] 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, 2, 6

[31] 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, 6, 7

[32] Mohammadreza Salehi, Niousha Sadjadi, Soroosh Baselizadeh, Mohammad H Rohban, and Hamid R Rabiee. Multiresolution knowledge distillation for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14902–14912, 2021. 2

[33] Shelly Sheynin, Sagie Benaim, and Lior Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8495–8504, 2021. 1, 2

[34] Jiaqi Tang, Hao Lu, Xiaogang Xu, Ruizheng Wu, Sixing Hu, Tong Zhang, Tsz Wa Cheng, Ming Ge, Ying-Cong Chen, and Fugee Tsung. An incremental unified framework for small defect inspection. In *European conference on computer vision*, pages 307–324. Springer, 2024. 2

[35] 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. 1, 6, 7

[36] Shenxing Wei, Xing Wei, Zhiheng Ma, Songlin Dong, Shaochen Zhang, and Yihong Gong. Few-shot online anomaly detection and segmentation. *Knowledge-Based Systems*, 300:112168, 2024. 3, 6, 7

[37] Guoyang Xie, Jinbao Wang, Jiaqi Liu, Feng Zheng, and Yaochu Jin. Pushing the limits of fewshot anomaly detection in industry vision: Graphcore. In *International Conference on Learning Representations*, 2023. 1, 2

[38] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8330–8339, 2021. 1, 2

[39] 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. 2

[40] Ximiao Zhang, Min Xu, and Xiuzhuang Zhou. Realnet: A feature selection network with realistic synthetic anomaly for anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16699–16708, 2024. 1, 6, 7

[41] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. In *International Conference on Learning Representations*, 2023. 1, 2, 6, 7

[42] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, pages 392–408. Springer, 2022. 6