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

Supplementary Material

A. Experimental Details

Backbone. We employ the ViT-L-14-336 backbone from the Open-CLIP implementation, resizing all images to a resolution of 518×518 prior to processing. All experiments are conducted on a single NVIDIA GeForce RTX 3090.

Text Prompt Branch. For the text prompt branch, we adopt the text template from April-GAN [2] and simplify it by replacing class names with the fixed terms "object" and "texture". We train using the Adam optimizer with a fixed learning rate of 0.005 for 5 epochs at a batch size of 16, and set the softmax temperature to 0.05. Features from the 12th, 16th, 20th, and 24th layers of the backbone—rich in semantic information—are utilized in this branch. When training on MVTec AD, we concatenate four images from the same category with a 20% probability to generate composite images that simulate multi-instance scenarios (as in April-GAN [2]). Additionally, images from the defect categories "misplaced" and "missing_cable" are discarded, since these often contain large abnormal regions outside the object that could mislead the model.

Patch-level Rarity Branch. In the patch-level rarity branch, features are extracted from the 6th, 12th, 18th, and 24th layers. Multi-scale features are obtained via 1×1 and 3×3 neighborhood average pooling to enhance visual representation. The rarity threshold X is set to 30%, and the maximum number of historical images $N_{I,\max}$ 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 (corresponding to the total patch count of three images), K is fixed at 3 for K-NN, and the loose degree Y is set to 1%.

Anomaly Detection. For pixel-level anomaly detection, we fuse the refined anomaly maps $\hat{\mathcal{A}}_{text}$ and $\hat{\mathcal{A}}_{rare}$ using weights that vary with the time step t. Specifically, we set:

$$\begin{cases} \lambda_{text} = 1, & \lambda_{rare} = 0, \quad t = 1, \\ \lambda_{text} = \frac{2}{3}, & \lambda_{rare} = \frac{2}{3}, \quad 1 < t \leq 4, \\ \lambda_{text} = \frac{1}{3}, & \lambda_{rare} = \frac{4}{3}, \quad t > 4, \end{cases} \tag{1}$$

which gradually increases the influence of the rarity branch as more test images are processed. We maintain the constraint $\lambda_{text} + \lambda_{rare}/2 = 1$ since the scale of $\hat{\mathcal{A}}_{rare}$ is roughly half that of $\hat{\mathcal{A}}_{text}$. The combined anomaly map is then upsampled to the original image resolution, and a Gaussian filter with $\sigma = 4$ is applied to smooth the final pixel-level result.

For image-level anomaly detection, we set the temperature parameter $\tau = 0.005$. The parameter B, which is used in the Image-level Re-scoring operation, is incremented by 1 every 20 tested images within the bounds of 0 and $B_{max} = 8$.

B. Loose Similarity Details

In RareCLIP, the patch-image similarity of a test patch is estimated by multiplying the similarity between the test patch and its nearest neighbor in the patch feature memory bank $\mathcal{M}_{\mathcal{F}}$ with the neighbor's patch-image similarity. Assuming that both the test patch and its nearest neighbor are similar to the same patch in a historical image, this estimation reduces to computing $\cos\theta_{AC}$ as $\cos\theta_{AB} \cdot \cos\theta_{BC}$, where $\theta_{AC} \in [0, \pi)$ is the angle between vectors A and C. Given the relation

$$|\theta_{AB} - \theta_{BC}| \leq \theta_{AC} \leq \min\big(\theta_{AB} + \theta_{BC},\, 2\pi - \theta_{AB} - \theta_{BC}\big),$$

we obtain the bound

$$\cos(\theta_{AB} + \theta_{BC}) \leq \cos\theta_{AC} \leq \cos(\theta_{AB} - \theta_{BC}).$$

Expanding further, the difference

$$\cos\theta_{AC} - \cos\theta_{AB} \cdot \cos\theta_{BC}$$

lies within

$$[-\sin\theta_{AB} \cdot \sin\theta_{BC},\, \sin\theta_{AB} \cdot \sin\theta_{BC}]\,,$$

yielding an error of $\pm\sqrt{(1 - \cos^2\theta_{AB})(1 - \cos^2\theta_{BC})}$. This error decreases as either $\cos\theta_{AB}$ or $\cos\theta_{BC}$ approaches 1—that is, when the test patch and its nearest neighbor (or the neighbor's patch-image similarity) are highly similar. In an industrial environment, normal patches typically find similar counterparts in other images, whereas abnormal patches do not. However, subtle differences among normal patches and variations in other regions prevent the similarities between normal patches from reaching 1, thereby gradually decreasing the estimated similarity of these normal patches after multiple multiplicative estimations. To address this, we loosen the requirement for similarities to reach 1, which has been shown to significantly boost RareCLIP's performance.

C. Anomaly Detection Visualization Results

We show the visualization of pixel-level anomaly detection results in Figure 1. Compared with other zero-shot methods [1–4, 7, 8], the proposed OnlineAD can better distinguish between normal regions and anomalous regions. For example, RareCLIP can easily find fine-grained anomaly in Cashew and component-missing anomaly in Pcb1.

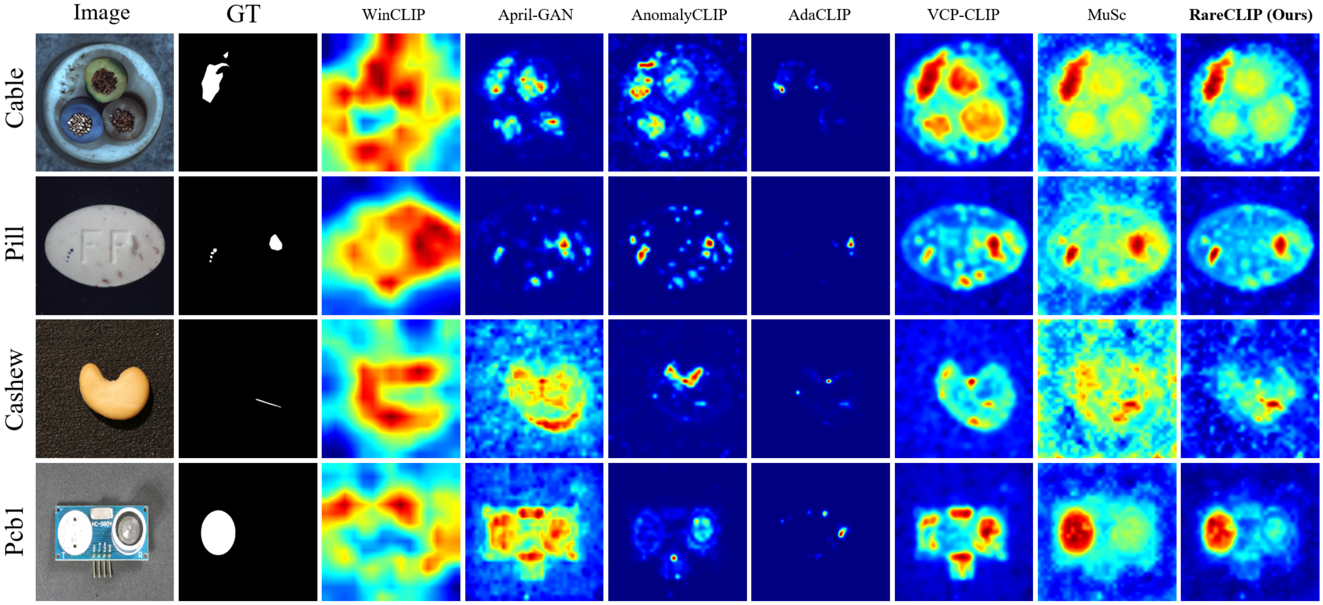

Figure 1. Visualization of pixel-level anomaly detection results on the MVTec AD (top two rows) and VisA (bottom two rows) datasets.

N_F	GPU (MB)	Time (ms)	MVTec AD		VisA	
			I-AUC	P-AUC	I-AUC	P-AUC
1369	4180	56.9	97.83	97.56	94.23	98.61
2738	4352	57.4	98.16	97.69	94.48	98.76
4107	4352	59.4	98.19	97.70	94.40	98.80
5476	4530	60.4	98.18	97.67	94.32	98.81
6845	4612	61.5	98.15	97.63	94.22	98.81

Table 1. Ablation study on the impact of different N_F values on the MVTec AD and VisA datasets. Here, N_F is set as an integer multiple of the patch number per image.

K	MVTec AD		VisA	
	I-AUC	P-AUC	I-AUC	P-AUC
1	98.17	97.52	94.03	98.78
2	98.19	97.64	94.26	98.80
3	98.19	97.70	94.40	98.80
4	98.12	97.73	94.44	98.78
5	98.08	97.76	94.44	98.77
6	98.04	97.77	94.41	98.76

Table 2. Ablation study on the effect of different K in K-NN.

α	GPU(MB)	Time(ms)	I-AUC	P-AUC
1/5	5244	91.4	93.18	98.83
1/3	6018	106.2	93.55	98.86
1/2	6942	124.9	93.66	98.86
1	9712	181.3	93.68	98.86

Table 3. Ablation study on the effect of different sampling ratio α in RareCLIP-d on the VisA dataset.

D. More Ablation Studies

Impact of Memory Size N_F. Table. 1 presents ablation study on the effect of varying N_F. As N_F increases, both the memory and time costs rise, while detection performance remains largely stable once $N_F \geq 2738$. This indicates that a moderate N_F can effectively reduce computational overhead without compromising performance.

Impact of K in KNN. Table. 2 shows the effect of varying the K value in the K-NN search. The results indicate that values of K that are either too low or too high lead to decreased performance, while a moderate K value yields the best results.

Sampling Ratio α. The ablation study of sampling ratio α in RareCLIP-d on VisA [9] dataset in Table. 3 shows that α = 1/3 achieves the best trade-off between computational costs and detection performance.

Hyperparameters in Anomaly Detection. As shown in Table. 4, we conduct an ablation study on hyperparameters $(\lambda_{text}, \lambda_{rare}, B, \tau)$ in Section 3.3 Anomaly Detection. The ablation results demonstrate that the performance is stable across various combinations of those hyperparameters.

Component Ablation. A component ablation study of RareCLIP is conducted in Table. 5 to show the impact of each component.

Results on BTAD. We evaluate RareCLIP on BTAD[6] without hyperparameter tuning and comparison results are shown in Table. 6.

$(\lambda_{text}, \lambda_{rare}, B, \tau)$	MVTec AD		VisA	
	I-AUC	P-AUC	I-AUC	P-AUC
(1/2, 2/2, 8, 1/200)	98.24	97.44	94.51	98.75
(2/5, 6/5, 8, 1/200)	98.24	97.62	94.54	98.79
(1/3, 4/3, 2, 1/200)	98.17	97.70	94.00	98.80
(1/3, 4/3, 5, 1/200)	98.19	97.70	94.34	98.80
(1/3, 4/3, 8, 1/100)	97.83	97.70	94.29	98.80
(1/3, 4/3, 8, 1/300)	98.26	97.70	94.00	98.80
(1/3, 4/3, 8, 1/200)†	98.19	97.70	94.40	98.80

Table 4. Ablation study on hyperparameters $(\lambda_{text}, \lambda_{rare}, B, \tau)$ related to anomaly score. † represents default.

TPB	IPRB			IRB	MVTec AD		VisA	
	PSM	SCS	LS		I-AUC	P-AUC	I-AUC	P-AUC
×	✓	✓	✓	✓	96.79	97.15	92.43	98.30
✓	×	✓	✓	✓	96.28	97.29	91.26	98.56
✓	✓	×	✓	✓	97.18	93.17	87.80	88.46
✓	✓	✓	×	✓	97.00	96.75	93.24	98.32
✓	✓	✓	✓	×	97.98	96.70	93.88	98.63
✓	✓	✓	✓	✓	98.19	97.70	94.40	98.80

Table 5. Component ablation study of RareCLIP.

Method	Mode	I-AUC	P-AUC
AdaCLIP [1]	offline	88.6	92.1
AnomalyCLIP [8]	offline	89.1	93.3
RareCLIP	offline	91.7	91.6
Musc* [4]	online-	94.6	97.4
RareCLIP	online	96.1	97.4

Table 6. Zero-shot results on the BTAD[6] dataset.

E. Online Few-shot Anomaly Detection

RareCLIP can be readily extended to the few-shot setting, where a limited number ($k = 1, 2, 4$) of normal images are available. Prior to testing, these k normal images are processed in online mode. The number of normal patch features N_k extracted from these k images is recorded in the patch feature memory bank, and a minimum threshold $N_{k,\text{min}}$ is enforced. If N_k falls below $N_{k,\text{min}}$ after applying SCS, the current N_k features are retained. For each test image, an additional anomaly map $\mathcal{A}_{few} \in \mathbb{R}^M$ is computed based on these N_k normal patch features:

$$\mathcal{A}_{few} = 1 - \text{LS}\left(\left\{ \max_{n\in[1,N_k]} \langle f_{test}, f^n_{\mathcal{M}} \rangle \;\middle|\; f_{test} \in \mathbf{F}_{test} \right\}\right), \tag{2}$$

where $f^n_{\mathcal{M}}, n \in [1, N_k]$ denotes the n-th normal patch feature in the memory bank and LS is the Loose Similarity operation to algin with IPRB. We then integrate $\mathcal{A}_{few}$ with $\mathcal{A}_{rare}$ using a weight parameter λ_k:

$$\mathcal{A}^*_{rare} = \frac{\left(1 - \frac{k}{t-1}\right) \mathcal{A}_{rare} + \lambda_k \, \mathcal{A}_{few}}{1 - \frac{k}{t-1} + \lambda_k}, \tag{3}$$

where the testing time step t starts from $k + 1$. We set $N_{k,\text{min}} = \frac{k}{4}M$ and $\lambda_k = \frac{k}{8}$ for $k = 1, 2, 4$, with M denoting the number of patches per image. A similar approach is applied to the Image-level Rarity Branch, where the locally aggregated image-level features F^{laif}_k from the k normal images are stored, and the maximum similarity between F^{laif}_{test} and F^{laif}_k is computed to refine c_{rare} with λ_k.

Table. 7 presents the comparative results of RareCLIP under different k-shot settings.

F. Limitations and Future Work

While RareCLIP achieves state-of-the-art performance in online zero-shot AD, it currently supports online detection only for a single category. This limitation restricts its direct application in industrial scenarios where multiple object categories must be monitored concurrently. In future work, we plan to extend RareCLIP to handle multi-category AD by exploring strategies such as shared and category-specific memory banks and adaptive mechanisms for processing diverse object types in a unified framework. In addition, further efforts will focus on reducing computational overhead and enhancing robustness in more complex real-world environments.

References

[1] 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. 1, 3

[2] 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, 4

[3] 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. 4

[4] 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, 3

[5] 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. 4

Dataset	Setting	Method	Mode	Image-level			Pixel-level			
				I-AUC	I-F1-max	I-AP	P-AUC	P-F1-max	P-AP	PRO
MVTec AD	1-shot	WinCLIP+	offline	93.1	93.7	96.5	95.2	55.9	-	87.1
		April-GAN	offline	92.0	92.4	95.8	95.1	54.2	51.8	90.6
		PromptAD	offline	94.6	-	-	95.9	-	-	87.9
		RareCLIP	offline	96.4	96.0	98.6	96.6	59.1	60.7	90.6
		RareCLIP	online	**98.5**	**98.0**	**99.5**	**97.9**	**65.0**	**67.2**	**93.6**
	2-shot	WinCLIP+	offline	94.4	94.4	97.0	96.0	58.4	-	88.4
		April-GAN	offline	92.4	92.6	96.0	95.5	55.9	53.4	91.3
		PromptAD	offline	95.7	-	-	96.2	-	-	88.5
		RareCLIP	offline	97.3	96.8	98.9	97.2	61.3	63.3	92.1
		RareCLIP	online	**98.6**	**98.1**	**99.5**	**98.0**	**65.7**	**67.9**	**93.8**
	4-shot	WinCLIP+	offline	95.2	94.7	97.3	96.2	59.5	-	89.0
		April-GAN	offline	92.8	92.8	96.3	95.9	56.9	54.5	91.8
		PromptAD	offline	96.6	-	-	96.5	-	-	90.5
		RareCLIP	offline	97.7	97.3	99.1	98.1	64.7	66.5	93.5
		RareCLIP	online	**98.7**	**98.1**	**99.6**	**98.2**	**66.5**	**69.0**	**94.0**
VisA	1-shot	WinCLIP+	offline	83.8	83.1	85.1	96.4	41.3	-	85.1
		April-GAN	offline	91.2	86.9	93.3	96.0	38.5	30.9	90.0
		PromptAD	offline	86.9	-	-	96.7	-	-	85.1
		RareCLIP	offline	92.6	88.6	94.2	98.1	44.6	39.0	92.1
		RareCLIP	online	**95.0**	**91.5**	**95.9**	**98.8**	**51.9**	**48.3**	**93.7**
	2-shot	WinCLIP+	offline	84.6	83.0	85.8	96.8	43.5	-	86.2
		April-GAN	offline	92.2	87.7	94.2	96.2	39.3	31.6	90.1
		PromptAD	offline	88.3	-	-	97.1	-	-	85.8
		RareCLIP	offline	93.4	89.4	94.9	98.4	47.0	41.6	92.9
		RareCLIP	online	**95.3**	**91.7**	**96.2**	**98.8**	**52.0**	**48.5**	**93.8**
	4-shot	WinCLIP+	offline	87.3	84.2	88.8	97.2	47.0	-	87.6
		April-GAN	offline	92.6	88.4	94.5	96.2	40.0	32.2	90.2
		PromptAD	offline	89.1	-	-	97.4	-	-	86.2
		RareCLIP	offline	94.6	90.5	95.5	98.8	49.6	45.4	93.7
		RareCLIP	online	**95.5**	**92.2**	**96.4**	**98.8**	**51.9**	**48.4**	**93.9**

Table 7. The comparison results with WinCLIP+ [3], April-GAN [2] and PromptAD [5] on the MVTec AD and VisA datasets under different k-shot settings.

[6] Pankaj Mishra, Riccardo Verk, Daniele Fornasier, Claudio Piciarelli, and Gian Luca Foresti. Vt-adl: A vision transformer network for image anomaly detection and localization. In *2021 IEEE 30th International Symposium on Industrial Electronics (ISIE)*, pages 01–06. IEEE, 2021. 2, 3

[7] 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. 1

[8] 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, 3

[9] 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. 2