

ADSeeker: A Knowledge-Grounded Reasoning Framework for Industry Anomaly Detection and Reasoning

Appendix

Anonymous CVPR submission

Paper ID

001

1. Dataset

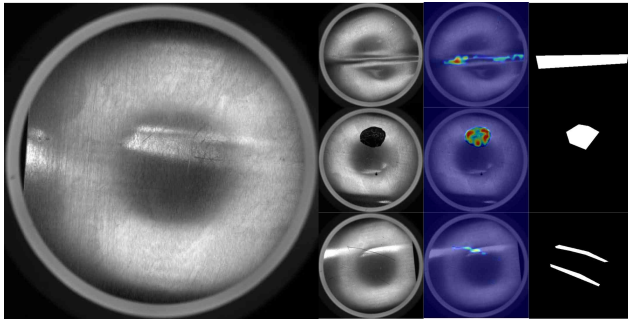


Figure 1. A synthesized presentation of some of the data from the Mula dataset.

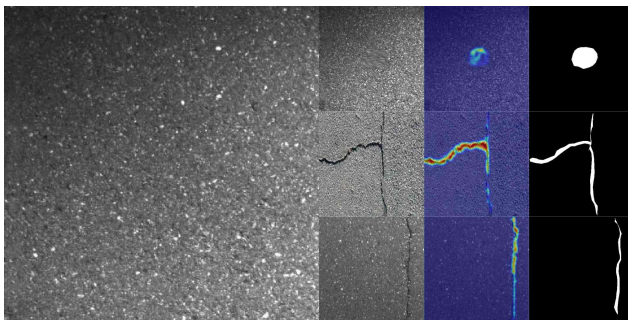


Figure 2. More partial data synthesis presentations of the Mula dataset.

We have built a new large-scale industrial anomaly detection dataset, Mula (Multi-Type Anomaly Dataset). In this dataset, each type of item contains perfect information at each scale, including sufficient normal samples, diverse anomaly samples, and accurate manually labeled masks. There are a total of 72 types of defects in Mula, which covers 26 industrial scenarios and contains data from a wide range of industrial domains. At the same time, to

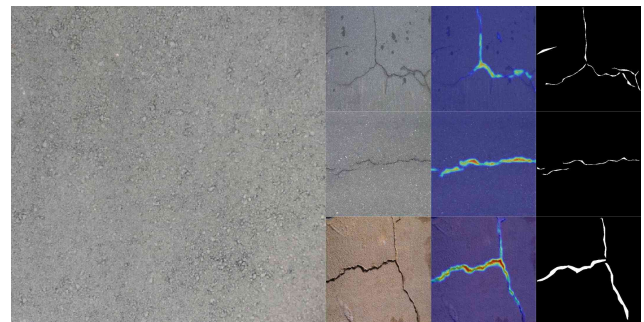


Figure 3. More partial data synthesis presentations of the Mula dataset.

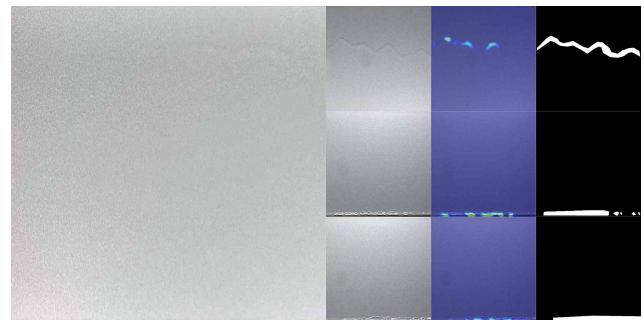


Figure 4. More partial data synthesis presentations of the Mula dataset.

reflect the actual complexity of industrial scene data, we have collected three types of data: RGB, grayscale, and X-ray, which corresponds to the specific realities of industrial scene data. The specific folder directory structure of Mula is exactly modeled after MVTec's structure, which makes it easy for other developers to use Mula's specific folder structure is modeled after MVTec's, making it easy for other developers to use our data sets directly. Also, using ADSeeker, we generated heat maps of the defect level of each anomaly sample. At the same time, using ADSeeker, we generated a

010
011
012
013
014
015
016
017
018
019

heat map of the defect level of each anomaly sample.

Most of the data was extracted from the factory of cooperative company, including images of actual processing scenarios. And we collected high-quality classes of MMIO for anomaly detection annotation. We organized six people to do data annotation work, and it took over 150 working hours. As for the intra-defect-type variability across products, Mula can be organized by defect type. Mula provides ADSeeker with type-level prior knowledge, which includes of 9 defect types. The same defect type contains at least 5 object types. To verify the consistent feature representation, we calculate the similarity between the image feature and the type-level text feature. Specifically, we calculate the similarity by $100.0 * \text{ImgF} * \text{TexF}$

As shown in Figures 1 through 4, we selected several types of items for display. In each figure, the leftmost is a normal sample, the three rows on the right correspond to three different types of anomalies, and the sub-figures on the right are, from left to right: anomalous original, anomalous degree heatmap, and anomalous mask. After this, we show a portion of normal samples for 26 types of items, as shown in Fig. 5. In addition, we show a portion of abnormal samples of each category, as shown in Fig. 6. We also show a part of the produced defect degree heat map, as shown in Fig. 7, and a part of the manually labeled mask map corresponding to the anomaly data, as shown in Fig. 8. Furthermore, we conducted a large number of anomaly detection experiments, the specific results of which are shown in Table 1.

our first goal is to present the Mula benchmark that can truly reflect the anomaly detection performance of models. The experimental data shows that test results of different models on the Mula are basically close to other AD benchmarks. The experiment follows the 0-shot setting and the test data set is not leaked during training, ensuring the fairness of the experiment. In order to expand the image dataset, our processing is mainly image composition. Considering that stitching will leave traces, we mainly use Poisson blending and local replacement to deal with inconspicuous small defects. Regarding geometric transformation, for texture images, we will intercept the normal part of the defective image and then perform interpolation and smoothing operations to return to the original size.

2. knowledge base

Our Agentic RAG (Retrieval Augmented Generation) framework has the ability to search from text to images, so that text information and image information can be searched accurately and quickly. We developed a knowledge base for the RAG framework that contains both text and image information. Based on the MVTec dataset and Visa, and with the help of the official description document of the dataset, we organized expert knowledge of various defects

with the corresponding case knowledge of the image samples to form a well-developed search information system. As shown in Figures 9 and 10, we present some samples of the knowledge base. We completed the knowledge base for each category in a certain order: production scenarios of the item, application scenarios of the item, and various anomalies of the item, including the names of the anomalies, the causes of the anomalies, and the possible consequences of the anomalies.

Regarding the LoRA fine-tuning dataset, we collected domain knowledge from MMAD and our knowledge base content. We manually created questions and used them for validity verification. The dataset contains most of the knowledge base content and is constructed using template examples. Compared with Anomaly-OV, ADSeeker uses the RAG architecture to replace the fine-tuning to solve the domain adaptation problem. Compared to the fine-tuned dataset of Anomaly-OV, ours is the first AD knowledge base for RAG, adding a dedicated domain knowledge.

3. Mula benchmark

To validate the significance of our Mula benchmark, we compare the performance of existing ZSAD approaches. As is illustrated in Table 1, our ADSeeker performs the best, with the average image-level AUROC and pixel-level AUROC reaching 97.63% and 97.5% , respectively. The IAD metrics obtained from Mula benchmark are basically commensurate with the inherent AD capabilities of the models which indicates that our benchmark can effectively capture the ZSAD performance of models and provide reliable and meaningful evaluations. Among the open-sourced ZSAD methods, earlier models like WinCLIP underperform compared to newer models like AnomalyCLIP, indicating that textual learnable prompt advancements in adaptive capabilities benefit performance in ZSAD tasks. It is worth noting that existing ZSAD methods perform poorly due to the difficulty of extracting type-level features from Mula, which should be taken into account in order to raise the ceiling for ZSAD tasks. During the test, we stopped utilizing Q2K RAG to avoid test dataset is leaked during training

4. Supplement to the model description

Seed4AD is trained on normal Mula samples. The advantage of ADSeeker is that no fine-tuning is required. ADSeeker-7B achieves the best reasoning capability. The AD Expert module can improve localization and discrimination capabilities, and semantic alignment can effectively lock abnormal locations, but feature enhancement will weaken some details of the image, resulting in a decrease in analysis and description capabilities. When using Q2K RAG to inject the knowledge base, the model's classification ability is significantly improved, but anomaly anal-

Model	Image-level			Pixel-level		
	I-Auroc	I-F1	I-AP	P-Auroc	P-F1	P-AP
ADSeeker	97.63	96.84	97.20	97.5	58.52	58.75
AdaCLIP	96.68	95.07	95.22	95.81	58.21	59.85
AnomalyCLIP	89.10	—	79.80	92.20	—	74.20
WinCLIP	78.88	62.45	—	80.29	14.69	—

Table 1. Performance comparison of ADSeeker and CLIP-based methods in MulA with the standard 0-shot setting. Anomaly Detection utilizes the average precision, AUROC and AUPRO to evaluate performance.

ysis is severely affected, which is related to the organization of the knowledge base.

[cla] is the general defect type. Given a defect query image i , it will be sent to Q2K RAG S to obtain the object category and defect category by clustering the encoded features [obj], [cla]= $S(i)$. The query image is used as input for HSP and generates anomaly pattern prompts injected into the encoder during the forward process. The vision encoder of CLIP performs forward inference on the query image to obtain the patch embedding, which is the mid-feature input of this module. As shown in L607-L620, the final feature is obtained by feature fusion. Compared with the vision encoder of Qwen-VL, we prefer to use the modality alignment capability of CLIP to predict anomalies through the embedding similarity of different modalities in the same feature space.

We verified the impact of RAG on reasoning performance. When ADSeeker did not use RAG, the reasoning ability of the model showed a significant decline. Compared with the baseline (Qwen-VL), although the information retrieval function is lost, the AD Expert uses fused feature to improve the model’s abnormal localization and detection capabilities, which are improved compared to the baseline.

5. Detailed Ablation Studies

To demonstrate the validity of each proposed module in ADSeeker, we conduct extensive ablation experiments by comparing the accuracy of ADSeeker in 2 subtasks among MMAD, the 2 subtasks are represented in our knowledge base. We primarily focus on 4 aspects: the AD-Expert module, the Q2K RAG module, the AD Expert module and the impact of LoRA fine-tuning on ADSeeker, our team constructed 43K Small-scale instruction tuning dataset. As is illustrated in the Table 3, a full-blooded version of ADSeeker performs best in these tasks.

For anomaly reasoning and analysis, it can be observed that each module has an irreplaceable role in outputs. ADSeeker could seek for best-match information in the SEEK M&V, to equip our model with domain knowledge, the

AD Expert module could provide prior knowledge related to anomaly localization and discrimination. The ability of trained models often exhibits a pattern of growth followed by a decline with the progression of training epochs. And training on such small-scale domain datasets is more likely to result in model overfitting and catastrophic forgetting of pre-trained knowledge. Furthermore, our Seek-Agent RAG framework outperforms LoRA fine-tuning in terms of generalizability and precision.

ADSeeker-7B achieves the best reasoning capability. The AD Expert module can improve Localization and Discrimination capabilities, and semantic alignment can effectively lock abnormal locations, but feature enhancement will weaken some details of the image, resulting in a decrease in analysis and description capabilities. When using Q2K RAG to inject the knowledge base, the model’s classification ability is significantly improved, but anomaly analysis is severely affected, which is related to the organization of the knowledge base. The fine-tuning effect has been pointed out in Sec5.3.

Dataset	I-AUROC	I-F1	I-AP	P-AUROC	P-F1	P-AP
Mvtec	94.12	95.20	97.91	96.17	57.95	61.14
headct	96.90	95.48	99.01	—	—	—
Isic	0	0	0	98.63	90.68	96.83
MPDD	89.87	91.90	93.97	96.53	47.73	46.48
ColonDB	—	—	—	95.06	69.92	76.98
brain-mri	97.19	96.84	99.47	—	—	—
ClinicDB	—	—	—	95.58	72.36	80.35
BTAD	93.69	96.46	99.11	98.59	69.94	73.13
MulA	96.34	95.75	96.20	97.63	60.02	60.05
Visa	91.97	89.92	95.22	97.14	42.21	35.90
Dtd	97.32	98.98	99.68	99.57	79.75	87.65

Table 2. ADSeeker’s anomaly detection results for each dataset.

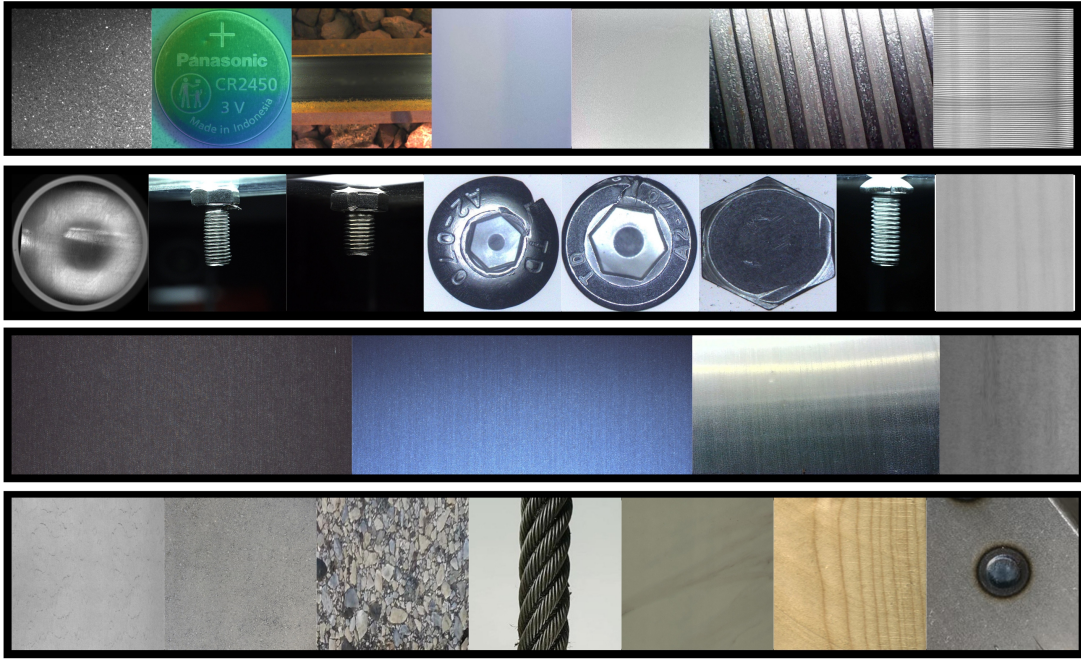


Figure 5. Presentation of normal samples of various items from the Mula dataset.

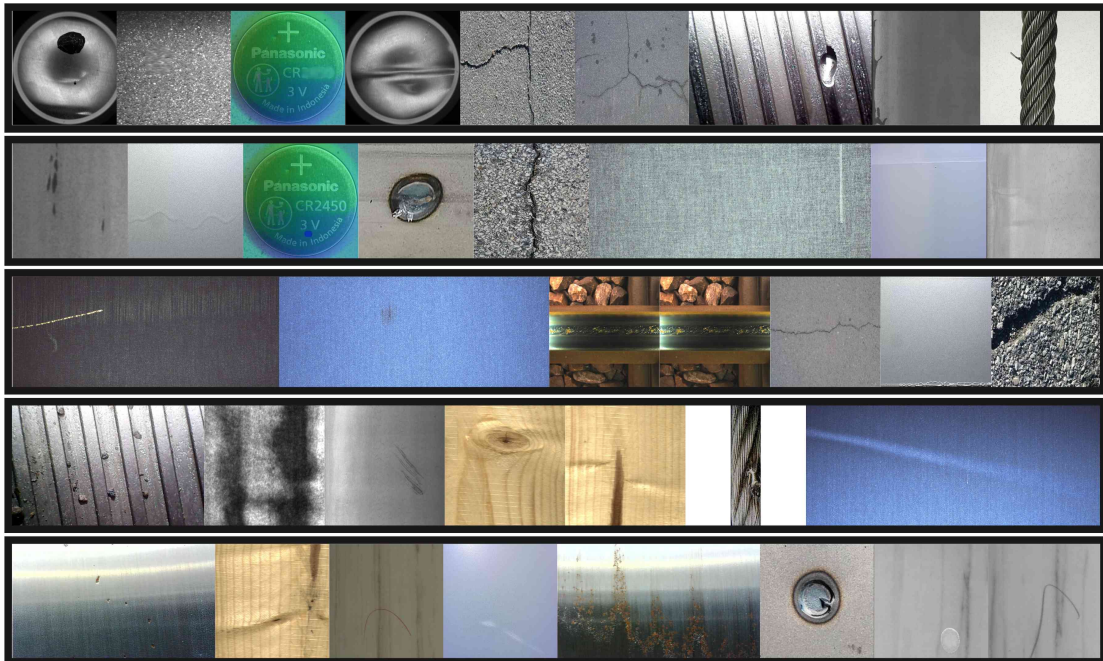


Figure 6. Presentation of anomalous samples of the parts of the Mula dataset for each type of item.

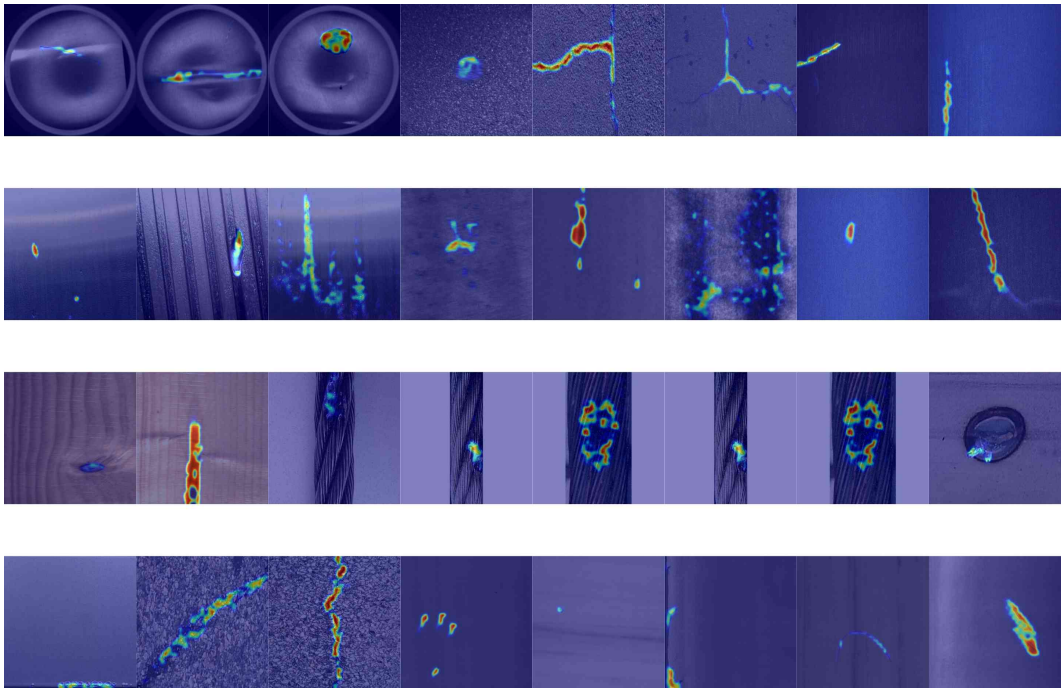


Figure 7. Presentation of selected information from the knowledge base.

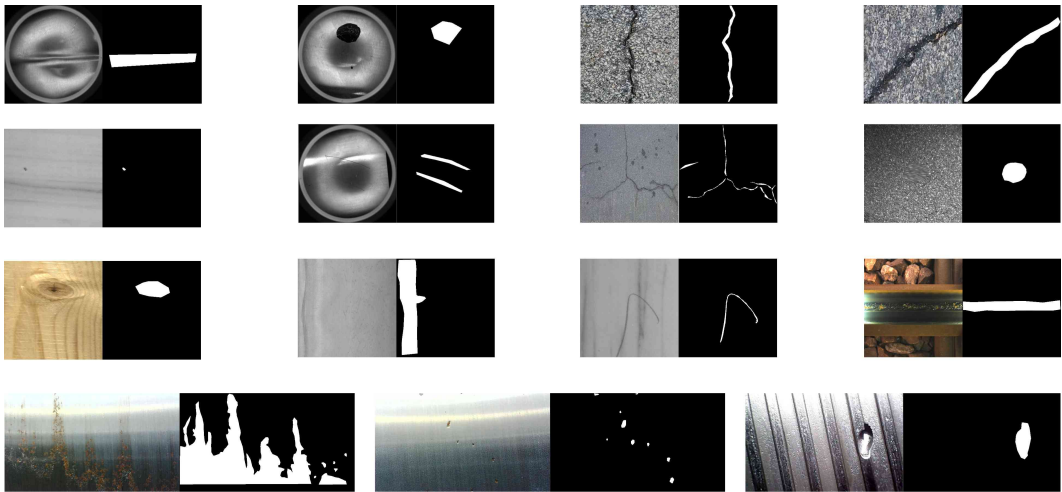


Figure 8. Presentation of selected information from the knowledge base.



Figure 9. Presentation of selected information from the knowledge base.

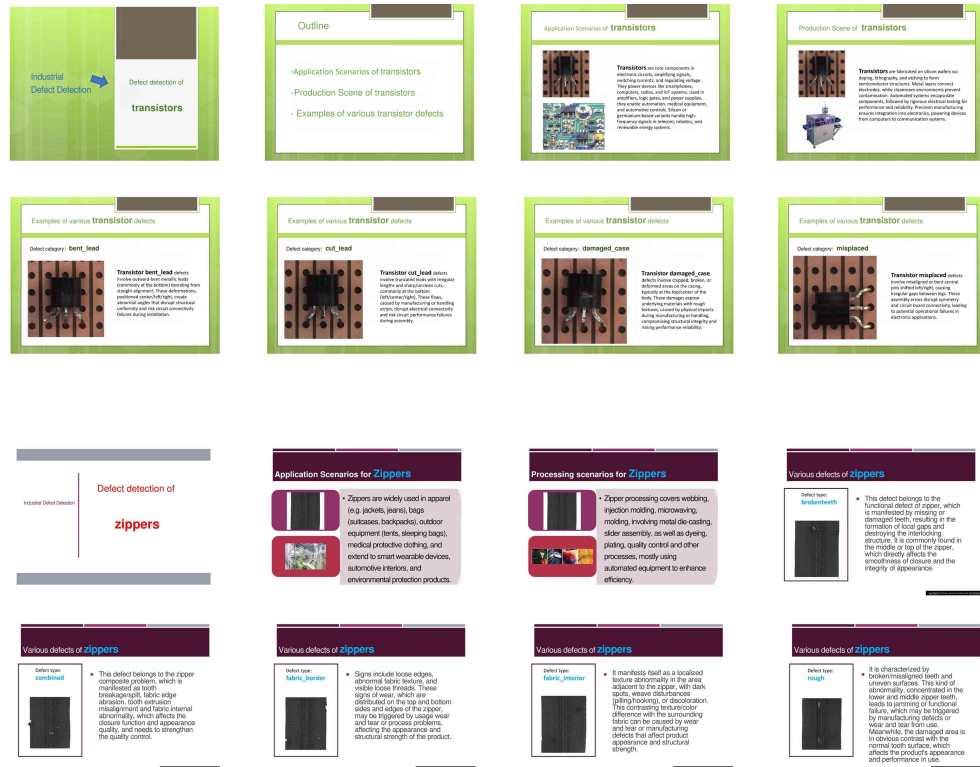


Figure 10. Presentation of selected information from the knowledge base.