Towards Training-free Anomaly Detection with Vision and Language Foundation Models

Supplementary Material

In this supplementary material, we provide details of experimental settings including data preprocessing and evaluation metrics. Additionally, quantitative results on MVTec LOCO [4], MVTec AD [3] and VisA [44] benchmarks are presented to demonstrate the effectiveness of our algorithm. Finally, we provide a comprehensive analysis and discussion of our framework.

7. Experimental Details

Data Preprocessing. Regarding vision and language foundation models including CLIP [28], DINOv2 [27] and SAM [17], we apply the same data preprocessing pipeline across MVTec LOCO, MVTec AD and VisA datasets to mitigate potential train-test discrepancy. Specifically, it involves channel-wise standardization with the pre-computed mean $[0.48145466, 0.4578275, 0.40821073]$ and standard deviation $[0.26862954, 0.26130258, 0.27577711]$ after normalizing each RGB image into $[0, 1]$, followed by bicubic interpolation based on the Pillow implementation.

Evaluation Metrics. Consistent with existing methods [3, 4], we report the results of the Area Under the Receiver Operator Curve (AUROC) documented in the body of the paper for the evaluation of image-level anomaly detection and pixel-level anomaly localization. Additionally, we supplement the F_1-max results in anomaly detection. The F_1-max score is computed from the precision and recall for the anomalous samples at the optimal threshold, which is a more straightforward metric to measure the upper bound of anomaly prediction performance across thresholds.

8. Quantitative Results

To elucidate the interaction between patch matching and composition matching in detecting logical and structural anomalies, we present the experimental results in Tab. 8, demonstrating that incorporating composition matching with patch matching improves results for logical AD and achieves comparable results for structural AD. Quantitative results indicate that the inclusion of composition matching significantly enhances detection performance for logical anomalies while maintaining comparable performance on structural anomalies. Additionally, we report the detailed subset-level results of LogSAD. Specifically, the results on MVTec LOCO [4] are presented in Tab. 9, and the results on MVTec AD [3] and VisA [44] benchmarks are depicted in Tab. 10 and Tab. 11, respectively.

Detectors	Structural	Logical	Average
Patch	87.3	67.6	77.4
Composition	58.0	78.2	68.1
Patch + Composition	85.8	82.0	83.9

Table 8. Image-level AUROC of multi-granularity detectors under 4-shot protocol on MVTec-LOCO dataset.

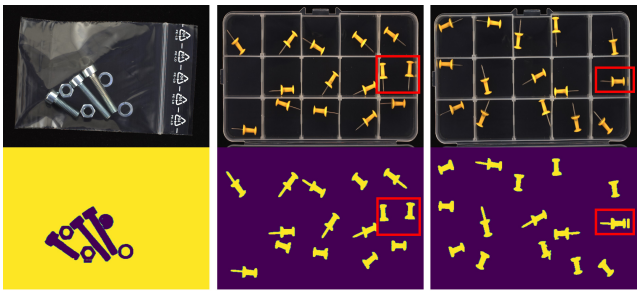

Figure 6. Failure cases of LogSAD.

9. Discussion

Canonical Normal Images in the MoT. The MVTec LOCO dataset [4] contains a varying number of normal subclasses in each category, *e.g.* 1, 3, 1, 1, 3 corresponding to "breakfast box", "juice bottle", "pushpins", "screw bag", and "splicing connectors". Specifically, the "juice bottle" category includes cherry, banana, and orange labels, while the "splicing connectors" category contains red, blue, and yellow cables. Thus, we sample 3 canonical normal images from the normal sub-classes in each category to maximize sub-class coverage in the training set, facilitating the establishment of comprehensive matching interests and compositional rules. We typically observe that the generated results from GPT-4V remain consistent across the provided normal images due to subtle visual variations within each sub-class. Notably, the sampled canonical normal images and GPT-4V are exclusively used for offline proposal generation, which operates independently of the anomaly detection algorithm.

In practice, the quality of generated proposals can be assessed in various perspectives, including the qualitative results of open-vocabulary semantic segmentation in terms of interests of thought, as depicted in Fig. 4, and quantitative results through interest matching and compositional matching, as shown in Tab. 6.

Failure cases. In addition, we present the failure cases in

Fig. 6 to address the limitations of our framework, including failures in open-vocabulary semantic segmentation and uncovered situations in compositional matching. For instance, (a) fails to distinguish the "hex nut" and the "ring washer"; (b) the number of pushpins is 15, but two pushpins appear in one division, which is not covered by matching rules; (c) fails in open-vocabulary semantic segmentation and counting due to the reflection of pushpins.

Computation Analysis. Previous methods, such as WinCLIP [15], PromptAD [20], AnomalyCLIP [43] fine-tuning with CLIP, and AnomalyGPT [12] fine-tuning with larger models (*e.g*. Vicuna-7B and Vicuna-13B), focus solely on structural AD but continue to struggle with logical AD. Our focus is on the training-free application of off-the-shelf foundation models for both logical and structural AD. Note that we use GPT-4V only for offline match proposal generation, and open-sourced foundation models including CLIP, DINOv2 and SAM are collaborated for anomaly detection with around 1.3B parameters. Consequently, Tab. 4 shows our method achieves state-of-the-art performance on structural AD datasets, outperforming training-based methods like PromptAD and AnomalyGPT. In addition, experimental results in Tab. 1, Tab. 2 and Tab. 3 demonstrate the effectiveness of our framework in both logical and structural AD.

Protocol	Breakfast Box		Juice Bottle		Pushpins		Screw Bag		Splicing Connectors		**Average**	
	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC
1-shot	85.0	88.0	85.6	78.1	75.7	78.0	80.7	70.6	78.2	77.7	81.0	78.5
2-shot	88.1	91.5	85.7	77.5	77.8	81.1	83.0	80.5	78.2	79.8	82.6	82.1
4-shot	89.9	94.4	88.2	84.3	81.4	82.5	84.1	81.5	84.7	88.6	85.7	86.3
full-data	92.0	95.7	94.0	95.2	81.3	83.6	85.2	83.2	91.3	93.5	88.8	90.2

Table 9. Image-level F_1-max and AUROC results on MVTec LOCO in few-shot and full-data protocols.

Category	image						pixel					
	1-shot		2-shot		4-shot		1-shot		2-shot		4-shot	
	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC
bottle	100	100	100	100	100	100	81.5	99.0	82.1	99.1	82.8	99.2
cable	88.0	90.4	87.9	91.2	87.8	90.8	60.7	96.7	62.7	97.4	63.1	97.6
capsule	94.0	92.0	94.9	92.7	97.3	94.1	50.3	98.0	50.5	98.3	51.7	98.4
carpet	98.9	99.4	98.9	99.3	98.9	99.4	67.6	99.2	67.4	99.2	67.5	99.2
grid	99.1	99.8	100	100	100	100	51.2	99.3	55.9	99.5	55.9	99.5
hazelnut	99.3	99.9	98.6	99.8	100	100	65.6	98.9	67.7	99.1	71.5	99.3
leather	99.5	99.9	99.5	99.9	100	100	49.4	99.3	48.0	99.3	48.9	99.4
metal_nut	99.5	99.6	99.5	99.7	100	100	74.7	96.0	76.7	96.3	84.6	97.8
pill	96.2	91.1	96.4	97.2	96.8	97.9	66.7	96.8	67.7	97.0	68.4	97.1
screw	92.2	92.4	92.2	92.4	92.2	92.4	18.7	95.6	22.3	96.6	27.5	97.4
tile	98.8	99.9	100	100	100	100	71.6	96.3	72.1	96.5	72.3	96.6
toothbrush	92.9	93.9	91.8	92.8	93.3	92.2	38.8	96.2	38.1	96.2	37.5	96.1
transistor	79.5	89.5	78.5	88.5	78.7	90.9	48.9	90.4	50.9	91.7	52.2	91.9
wood	99.2	99.8	99.2	99.7	99.2	99.8	70.2	97.0	70.2	97.0	70.2	97.0
zipper	96.8	93.5	98.3	94.7	99.2	97.6	56.1	96.6	58.2	97.1	58.6	97.3
average	95.6	96.1	95.7	96.5	96.2	97.0	58.1	97.0	59.4	97.3	60.8	97.6

Table 10. Image-level/pixel-level F_1-max and AUROC results on MVTec AD.

Category	image						pixel					
	1-shot		2-shot		4-shot		1-shot		2-shot		4-shot	
	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC	F_1-max	AUROC
candle	88	92.5	86.7	92.0	87.4	92.4	36.2	98.2	36.5	98.9	36.2	98.9
capsules	91.9	96.0	92.5	96.5	93.1	97.1	40.3	96.6	42.0	96.8	44.9	97.8
cashew	81.2	78.2	84.2	83.5	91.5	93.7	62.8	98.5	63.0	98.6	62.8	98.5
chewinggum	95.0	97.7	96.0	97.3	97.5	98.7	69.6	99.5	70.1	99.6	69.4	99.5
fryum	91.5	93.7	92.8	96.0	96.5	98.3	33.7	93.9	38.7	95.0	41.8	95.1
macaroni1	84.7	89.6	85.1	90.9	90.1	93.7	27.1	98.6	29.5	98.9	29.0	99.1
macaroni2	69.2	68.3	68.7	68.5	72.7	75.9	14.4	97.9	13.2	98.1	16.9	98.3
pcb1	85.3	91.3	85.1	91.8	84.3	91.3	52.0	98.0	48.7	98.2	48.5	98.2
pcb2	79.1	84.6	80.5	86.5	80.4	87.3	36.1	98.0	36.7	98.0	37.8	98.2
pcb3	76.2	82.3	81.9	87.6	87.9	93.5	40.6	97.6	38.8	98.1	40.4	98.5
pcb4	82.6	84.9	85.5	89.3	89.9	94.7	39.1	94.8	40.8	94.5	43.6	96.0
pipe_fryum	98.0	99.5	98.0	99.7	98.0	99.5	59.2	99.1	58.2	99.1	60.3	99.2
average	85.2	88.2	86.4	90.0	89.1	93.0	42.6	97.6	43.0	97.8	44.3	98.1

Table 11. Image-level/pixel-level F_1-max and AUROC results on VisA.