Towards Better Zero-Shot Anomaly Detection under Distribution Shift with CLIP

Jiyao Gao, Chengxin He, Lei Duan and Jie Zuo*

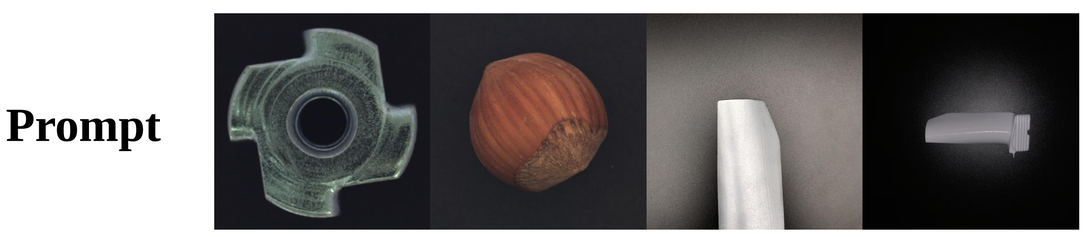
zuojie@scu.edu.cn

Sichuan University

Background

Industrial Anomaly Detection (IAD) is a critical task in real-world computer vision applications, focusing on identifying defective products during quality control. Traditionally, IAD relies on unsupervised methods due to the scarcity of anomalous samples. Zero-Shot Anomaly Detection (ZSAD) has recently gained traction, leveraging large-scale Vision Language Models (VLMs) like CLIP, showing impressive zero-shot classification abilities across different domains.

The existing VLM-based methods try to manufacture appropriate prompts for the VLMs under IAD scenario [1]. This solution works well when the test and training samples are from the same distribution. However, a more realistic scenario is that the test samples are not guaranteed to be from the same distribution, which is called distribution shift. This phenomenon can be introduced by many environmental conditions changes, such as shooting angles, lighting and background [5]. Moreover, the distribution shift is arbitrary and unpredictable. Thus, we raised this problem: *Without expert-designed prompts beforehand, how to make the VLMs still perform well under variable IAD scenarios?*



Method

[2] had shown that the text vector produced by CLIP's text encoder can be treated as training samples for FNN as a distribution shift robust classifier. [3] further investigated the randomly generated text vector can be used in IAD. Based on these findings, we believe that by generating more diverse text vectors, the FNN can be more robust to distribution shifting samples. Meanwhile, we must ensure that these text vectors will not deviate from their original meanings, i.e., "normal" and "abnormal".

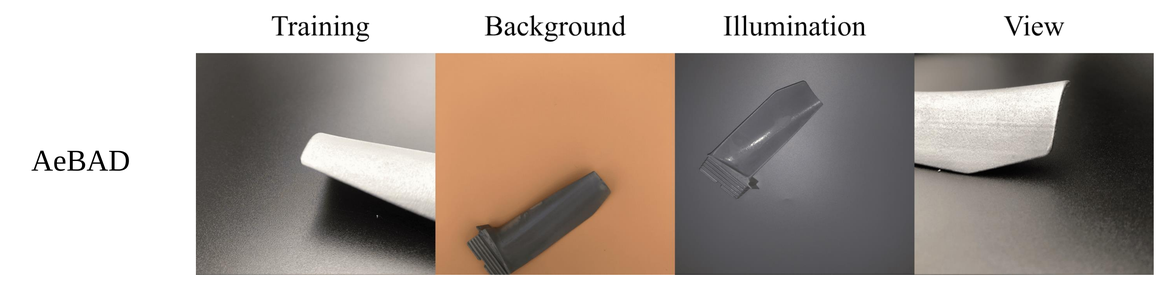
More specifically, our proposed text vectors is consisting of three parts: semantic words, distribution words and diversity words.

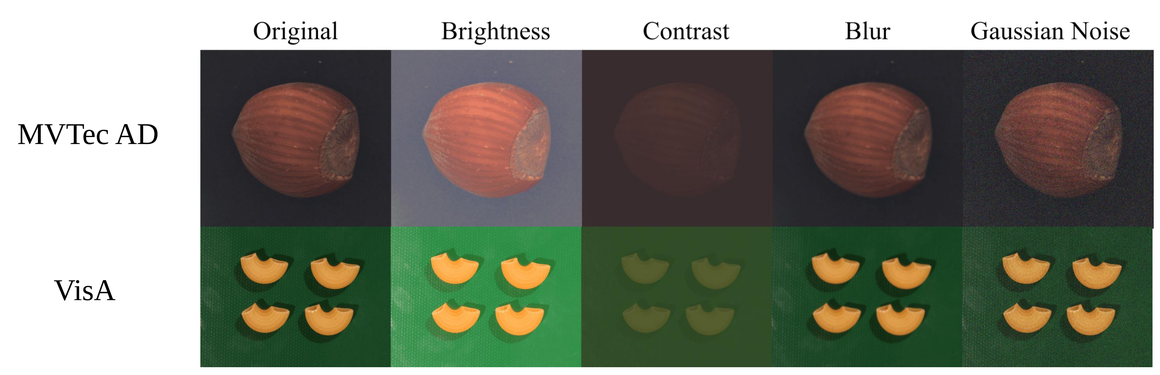
The diversity words and distribution words are learnable text vectors, while the semantic words are human-readable texts. As mentioned before, our goal is to obtain as many diverse samples, while maintaining their normal and abnormal semantics. Thus, we incorporate the three parts to achieve these two goals simultaneously. The two kinds of semantic words contain the class name of samples to be examined and an adjective, for an example, "broken" or "damaged", to denote normal and abnormal objects. These words are frozen, which ensures the generated vectors contain the most important normal/abnormal semantics. Oppositely, the diversity words are expected to be different form each other. We try to minimize the cosine similarity between them to achieve this. As for the distribution words, we provide a mixture guidance, which combined knowledge and diverse loss.

The optimized distribution and diversity words are combined with normal and anomaly semantic words to form training samples for a feed-forward neural network (FNN). This FNN is trained to perform anomaly detection, achieving robustness against distribution shifts without using any actual images.

Experiments

In this paper, we use three real-world and synthetic anomaly detection datasets, AeBAD , MVTec AD and VisA. While AeBAD is a To better simulate different kinds of distribution shift in real world, we follow [4] to apply four kinds of image corruptions to the test samples.

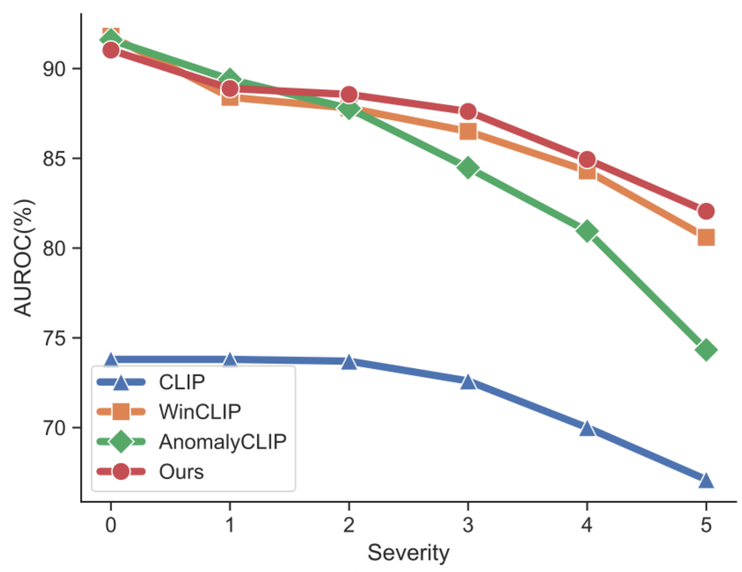




Our method achieves remarkable results on these datasets:

Methods	AeBAD	MVTec						VisA					
		brightness	contrast	blur	noise	mean	none	brightness	contrast	blur	noise	mean	none
MAEDAY	57.0	60.3	69.2	61.9	60.2	62.9	60.1	53.2	57.9	55.3	58.3	56.2	54.4
ACR	-	84.5	82.7	83.0	83.2	83.3	85.8	67.3	71.8	72.0	70.9	70.5	73.5
CLIP	53.1	74.1	68.1	74.2	74.2	72.7	73.8	58.3	58.1	58.3	58.2	58.2	58.2
CLIP-AC	53.0	73.2	69.0	73.7	72.7	72.2	73.9	59.4	57.6	57.8	59.6	58.6	58.5
RWDA	74.0	85.2	83.2	87.7	85.8	85.5	91.0	68.8	73.1	73.1	73.3	72.1	78.1
WinCLIP	71.9	85.7	85.9	86.9	**87.6**	86.5	**91.8**	67.9	72.2	73.3	74.3	71.9	78.1
AnomalyCLIP	73.0	84.6	86.2	86.1	81.0	84.5	91.5	**76.9**	75.0	**79.7**	77.7	**77.3**	**81.9**
Ours	**80.7**	**86.3**	**88.0**	**88.5**	**87.6**	**87.6**	91.0	72.5	**75.9**	76.1	**78.4**	75.7	78.8

Our method is also robust to different severity of corruptions:



References

[1] Jeong, Jongheon, et al. "Winclip: Zero-/few-shot anomaly classification and segmentation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[2] Cho, Junhyeong et al. "PromptStyler: Prompt-driven Style Generation for Source-free Domain Generalization." 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2023): 15656-15666.

[3] Tamura, Masato. "Random Word Data Augmentation with CLIP for Zero-Shot Anomaly Detection." British Machine Vision Conference (2023).

[4] Cao, Tri Thien et al. "Anomaly Detection under Distribution Shift." 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2023): 6488-6500.

[5] Zhang, Zilong, et al. "Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction." Computers in Industry 151 (2023): 103990.

SICHUAN UNIVERSITY