





# Zero-Shot Anomaly Detection with Pre-trained Segmentation Models

Matthew Baugh<sup>1</sup>, James Batten<sup>1</sup>, Johanna P. Müller<sup>2</sup>, and Bernhard Kainz<sup>1,2</sup>

<sup>1</sup> Imperial College London, UK

`matthew.baugh17@imperial.ac.uk`

<sup>2</sup> Friedrich–Alexander University Erlangen–Nürnberg, DE

**Abstract.** This technical report outlines our submission to the zero-shot track of the Visual Anomaly and Novelty Detection (VAND) 2023 Challenge. Building on the performance of the WinCLIP framework, we aim to enhance the system’s localization capabilities by integrating zero-shot segmentation models. In addition, we perform foreground instance segmentation which enables the model to focus on the relevant parts of the image, thus allowing the models to better identify small or subtle deviations. Our pipeline requires no external data or information, allowing for it to be directly applied to new datasets. Our team (Variance Vigilance Vanguard) ranked third in the zero-shot track of the VAND challenge, and achieve an average F1-max score of 81.5/24.2 at a sample/pixel level on the VisA dataset.

## 1 Introduction

The creation of dedicated datasets for industrial anomaly detection has led to a heightened interest in the area, and with it huge progress. Because of this, the state of the art has advanced so far that unsupervised industrial anomaly detection appears close to solved, with new methods often reporting an Area Under the Receiver Operating Characteristic (AUROC) above 98.0 at both a sample and pixel level [5,8,9] on the MVTec dataset [1]. However, for these methods to perform well they often require a high number of normal samples to train on, making such algorithms difficult to apply in a real-world setting. There is therefore a need for more data-efficient methods, able to identify anomalies when trained on only a few normal samples. WinCLIP[3] pushed this idea further, proposing a zero-shot method anomaly detection and localisation, leveraging language guidance to provide a signal for normality in the absence of normal images.

We adapt WinCLIP by incorporating zero-shot segmentation models to better localise anomalies. We also use foreground segmentation to focus our model on each instance of the object within the image, leading to better segmenting of smaller anomalies. Through this, our method achieves an average F1-max score of 24.2 at a pixel-level on the VisA dataset[10] and ranked third on the zero-shot track of the Visual Anomaly and Novelty Detection (VAND) 2023 Challenge.

## 2 Method

Our method maintains the core principles of WinCLIP [3] by using CLIP-based models to identify anomalous examples, but in order to better localise the anomalies we use a combination of zero-shot segmentation models. Our pipeline (Fig. 1) can be broken down into foreground extraction, image tiling, prompt generation, prediction (at both a tile and pixel level) and finally prediction aggregation.

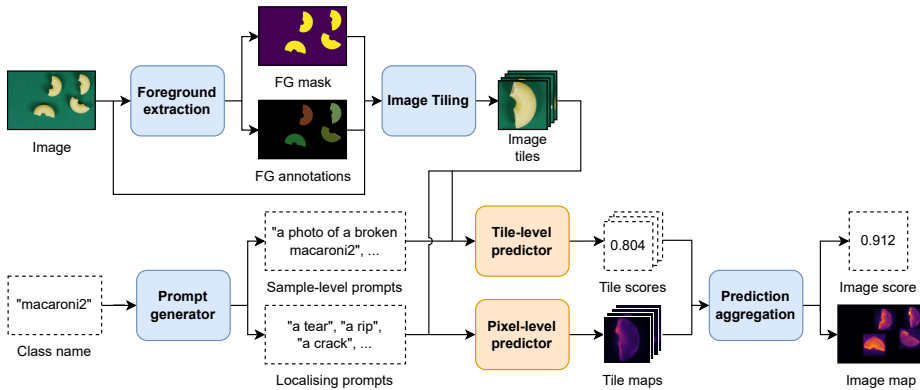


Fig. 1: Our zero-shot anomaly detection pipeline.

**Foreground extraction:** To identify the foreground of the image we combine dichotomous image segmentation [7] with SegmentAnything (SAM) [4], using the dichotomous segmentation to filter the annotations produced by SAM to identify which form the foreground of the image. These annotations are then combined to produce the final foreground mask. We found that this performs better than directly using the dichotomous segmentation as it often overestimates the size of an object to include part of the shadows. As SAM normally produced a separate annotation for the object and its shadow, our simple filtering (requiring 80% of the mask to be covered by the dichotomous segmentation) mitigated most of these cases.

**Image tiling:** The primary purpose of our image tiling was to divide the image up into each instance of the object. To do this we crop a square centred on each connected component of the foreground mask with a minimum resolution of  $352 \times 352$ , which is the maximum input size of our later models, as we found that using tiles smaller than this degraded performance as the images would be distorted as part of the model’s preprocessing steps. An exception to this was made for components with a higher aspect ratio ( $> 1.5$  in either axis) which SAM identified as having many constituent parts ( $> 20$ ). For these objects taking a square bounding box would mean that the background would take up a high proportion of the image, making it difficult for the segmentation models to notice deviations in the components of the object. In these cases we tile along the long

axis of the bounding box of the object, taking steps of half the short axis length, to ensure that all parts of the object are included in the centre of at least one tile.

**Prompt generator:** We used WinCLIP’s [3] compositional prompt ensemble for the sample-level prompts, only extending the list of normal and abnormal states in order to increase robustness (Fig. 2 a, b). For the localising prompts we use a list of generic nouns that describe an anomalous region of an object (Fig. 2 c). Such prompts were more suited than directly reusing the WinCLIP prompts to localise the defects, as the WinCLIP prompts describe the entire image so would produce much broader segmentations than those just describing the anomaly. The same list of localising prompts is used for all classes.

**Tile-level predictor:** To compute an anomaly score for each tile we take a similar approach to WinCLIP [3], comparing the CLIP embeddings of the different sample-level prompts to the embedding of each image tile. However, rather than averaging themselves, we average the alignment between each prompt and the tile image embedding. We favour this as although you would expect the embeddings of the “normal” prompts to form a single cluster, as they are describing the same concept of a normal object, the anomalous prompts are likely to be spread a lot more sparsely as being anomalous covers a wide variety of concepts. This means that taking the average embedding of the anomalous prompts is not as meaningful, but comparing the average cosine similarities avoids this issue.

**Pixel-level predictor:** For pixel-level predictions, for each tile we use CLIPSeg [6] to produce a segmentation from each of the localising prompts. We then use a harmonic average across the prompt segmentations to focus on the regions which consistently give a higher activation.

**Prediction aggregation:** As the tile-level predictor is generally more accurate and robust than the pixel-level predictor, we scale the pixel-wise predictions for each tile by its corresponding tile-level prediction. These tiles are then re-arranged into the original image space, with pixels belonging to multiple tiles being averaged across the predictions. To compute the sample-level prediction, we first average the tile-level predictions across each foreground component, as different tiles of the same foreground are often heterogeneous so have different distributions of tile-level scores. We then take the average of the top 25% of foreground component scores, which is particularly useful in the multi-instance cases as it avoids some of the noise present in the scoring of the normal foreground regions.

### 3 Results

Following WinCLIP [3] we used the F1-score at the optimal threshold (F1-max) to assess our method as it is less influenced by class imbalance, which is particularly prevalent in the segmentation evaluation as the anomalies are often quite small relative to the size of the image. We provide our F1-max at both a sample level (Tab. 1) and pixel level (Tab. 2) on the VisA dataset [10], comparing to WinCLIP [3] as a baseline while also including the results of the VAND challenge

winner APRIL-GAN [2]. Our use of additional sample-level prompts and averaging the cosine similarities achieves a new state-of-the-art for sample-wise F1-max (81.5), while our pixel-wise performance greatly improves over the baseline.

	pcb1	pcb2	pcb3	pcb4	capsules	candle	macaroni1	macaroni2	cashew	chewinggum	fryum	pipe_fryum	Mean
WinCLIP	71.0	67.1	<b>71.0</b>	74.9	83.9	<b>89.4</b>	74.2	69.8	<b>88.4</b>	<b>94.8</b>	82.7	80.7	79.0
APRIL-GAN	66.9	<b>70.1</b>	66.7	<b>87.3</b>	77.6	77.8	71.1	69.1	84.8	93.7	<b>91.7</b>	87.7	78.7
Variance Vigilance Vanguard	<b>74.3</b>	67.1	70.2	<b>87.3</b>	<b>84.9</b>	82.1	<b>83.3</b>	<b>76.9</b>	82.3	94.4	84.8	<b>90.0</b>	<b>81.5</b>

Table 1: Sample-wise results, F1-max compared with baseline WinCLIP and challenge winner APRIL-GAN[2]

	pcb1	pcb2	pcb3	pcb4	capsules	candle	macaroni1	macaroni2	cashew	chewinggum	fryum	pipe_fryum	Mean
WinCLIP	2.4	4.7	10.3	32.0	9.2	22.5	7.0	1.0	13.2	41.1	22.1	12.3	14.8
APRIL-GAN	12.5	23.4	21.7	31.3	48.5	39.4	35.5	13.7	22.9	78.5	29.7	30.4	32.3
Variance Vigilance Vanguard	29.5	11.0	4.7	21.7	31.9	20.2	24.6	7.2	24.5	63.4	31.3	19.6	24.2

Table 2: Pixel-wise results, F1-max compared with baseline WinCLIP and challenge winner APRIL-GAN[2]

## 4 Conclusion

We have greatly improved the segmentation ability of WinCLIP by incorporating zero-shot segmentation models. However, there is certainly more scope for improvement in the localising ability, as many of the models we use struggle due to the magnitude of the domain shift from their original testing data to that of industrial anomaly detection. This problem was amplified by many of the anomalies being exceedingly small and subtle. As foundation models continue to progress we are excited to see how their better representations can be leveraged to better solve the task of zero-shot anomaly detection. At a sample level, our results improve incrementally over WinCLIP [3], but there is still much work to be done to elevate zero-shot anomaly detection to be closer to the performance of unsupervised models.

## References

1. Bergmann, P., Batzner, K., Fauser, M., Sattlegger, D., Steger, C.: The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision* **129**(4), 1038–1059 (2021)
2. Chen, X., Han, Y., Zhang, J.: 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)

3. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: CVPR 2023 (2023), <https://www.amazon.science/publications/winclip-zero-few-shot-anomaly-classification-and-segmentation>
4. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
5. Liu, Z., Zhou, Y., Xu, Y., Wang, Z.: Simplenet: A simple network for image anomaly detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20402–20411 (June 2023)
6. Lüddecke, T., Ecker, A.: Image segmentation using text and image prompts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7086–7096 (2022)
7. Qin, X., Dai, H., Hu, X., Fan, D.P., Shao, L., Gool, L.V.: Highly accurate dichotomous image segmentation. In: ECCV (2022)
8. Roth, K., Pemula, L., Zepeda, J., Schölkopf, B., Brox, T., Gehler, P.: Towards total recall in industrial anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14318–14328 (June 2022)
9. Tien, T.D., Nguyen, A.T., Tran, N.H., Huy, T.D., Duong, S.T., Nguyen, C.D.T., Truong, S.Q.H.: Revisiting reverse distillation for anomaly detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 24511–24520 (June 2023)
10. Zou, Y., Jeong, J., Pemula, L., Zhang, D., Dabeer, O.: Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 392–408. Springer Nature Switzerland, Cham (2022)

## 5 Appendix

(a) Additional normal state-level prompts	(b) Additional anomalous state-level prompts	(b) Localising prompts
– "good [o]"	– "broken [o]"	– "a tear"
– "normal [o]"	– "bad [o]"	– "a rip"
– "amazing [o]"	– "flawed [o]"	– "some damage"
– "pristine [o]"	– "defective [o]"	– "a fault"
– "undamaged [o]"	– "[o] in poor condition"	– "a break"
– "[o] in good condition"	– "worn [o]"	– "an abnormality"
– "unbroken [o]"	– "[o] with scratches"	– "a defect"
– "[o] without any imperfections"	– "[o] with marks"	– "a crack"
– "[o] without any scratches"	– "[o] with imperfections"	– "an anomaly"
– "[o] without any marks"	– "cracked [o]"	– "a missing component"
– "complete [o]"	– "faulty [o]"	– "an error"
– "new [o]"	– "incomplete [o]"	– "a mark"
	– "bent [o]"	– "a cut"
	– "snapped [o]"	– "a dent"
	– "scratched [o]"	– "a scratch"
	– "shattered [o]"	– "an imperfection"
	– "fractured [o]"	– "a blemish"
	– "burst [o]"	– "a mistake"
	– "[o] in pieces"	– "an error"

Fig. 2: Lists of prompts used in our pipeline, excluding those from the original WinCLIP [3].