263-3300-10L Data Science Lab: Few-Shot Anomaly Detection in a Real-World Aero-Engine Blade Inspection Use Case

Berke Arda*
bearda@student.ethz.ch
Swiss Federal Institute of Technology Zurich
(ETH Zurich), Switzerland

Katarína Osvaldová*
kosvaldova@student.ethz.ch
Swiss Federal Institute of Technology Zurich
(ETH Zurich), Switzerland

Florian Scheidegger
IBM Research Europe - Zurich, Switzerland

Julia Vogt
julia.vogt@inf.ethz.ch
ETH AI Center, ETH Zurich, Switzerland

Abstract

The aim of this study is an evaluation of state-of-the-art few-shot anomaly classification methods in a real-world aero-engine blade inspection use case using the AeBAD dataset. The study focuses on WinCLIP and its extended variant, WinCLIP+, under challenging domain shifts that involve changes in illumination, background, and viewpoint.

Our results show that even in the zero-shot variant, WinCLIP performs reasonably well, while WinCLIP+ provided with a small number of normal samples offers further improvement.

A fair comparison with a benchmark method trained on different amounts of data reveals the robustness of WinCLIP+ in low-data regimes, narrowing the performance gap. This study highlights the potential of few-shot learning for real-world anomaly detection tasks with limited labelled data.

CCS Concepts

• **Computing methodologies** → **Anomaly detection**.

Keywords

Anomaly Detection, Few-Shot Learning, WinCLIP, AeBAD, Domain Shift, Computer Vision, Low-Data Regimes

1 Introduction

Anomaly detection plays an essential role in industrial inspection, where the identification of defects in components is essential to ensure quality assurance and operational reliability. Traditional methods for anomaly detection often rely on large labelled datasets for training, which are costly and time-consuming to acquire. In real-world scenarios, such as aero-engine blade inspection, labelled anomalies are particularly scarce, creating the need for methods that can generalize well with limited examples.

Few-shot anomaly detection addresses this challenge by leveraging a small number of labelled samples, enabling effective model training in low-data regimes. However, achieving robust performance under domain-shifts, such as changes of the background, illumination or perspective, remains a significant hurdle. Models must not only identify anomalies accurately but also generalize across diverse operational conditions.

In this project, we investigate the effectiveness of state-of-the-art methods for few-shot anomaly detection in the context of aero-engine blade inspection using the AeBAD dataset. This dataset, by design, contains domain shifts that mirror real-world variations, making it an ideal benchmark for evaluating model robustness.

We implement and evaluate WinCLIP, a zero-shot anomaly detection method, with its enhanced few-shot version, WinCLIP+ [3], while ensuring fair comparisons against existing benchmarks, such as MMR [12], under reduced-data conditions.

This report is structured as follows: Section 2 discusses related work in anomaly detection and few-shot learning. Section 3 provides an overview of the AeBAD dataset, including the domain-shift challenges it presents. Section 4 explains our process in more detail, followed by Section 5, which presents the results and their analysis. Section 6 discusses the time vs. performance trade-off, and Sections 7 and 8 conclude the report with key findings and future research directions.

2 Related Work

Anomaly detection has been a long-standing challenge in computer vision, particularly in industrial applications where detecting rare and diverse defects is critical to quality assurance. Early methods predominantly relied on handcrafted features and traditional machine learning techniques, such as Support Vector Machines (SVMs) [8] and Principal Component Analysis (PCA) [4]. These approaches, while effective for structured data, often struggled with complex visual patterns and high-dimensional image data.

*Equal contribution.

263-3300-10L Data Science Lab, December 20, 2024, Zurich, Switzerland

Author contributions using the CRediT framework [1]:
Berke Arda: Methodology, Software, Investigation, Writing - Original Draft
Katarína Osvaldová: Methodology, Software, Visualization, Investigation, Writing - Original Draft
Florian Scheidegger: Conceptualization, Supervision
Julia Vogt: Supervision, Project Administration

The advent of deep learning has significantly advanced anomaly detection capabilities. Reconstruction-based methods, such as Autoencoders [2] and Generative Adversarial Networks (GANs) [7], have become popular due to their ability to learn rich representations of normal data. However, these methods are sensitive to domain shifts, such as changes in illumination or background, which often occur in real-world industrial scenarios.

Few-shot learning has emerged as a promising paradigm for addressing the data scarcity challenge in anomaly detection. Methods like Prototypical Networks [9] and Relation Networks [10] focus on learning representations from minimal examples. More recently, self-supervised approaches, such as PatchCore [6] and DRAEM [11], have shown promise in leveraging unsupervised pre-training for few-shot anomaly detection. Despite these advancements, existing methods often require extensive fine-tuning or struggle with generalization under domain shifts.

Our work builds on these approaches by evaluating WinCLIP, a zero-shot anomaly detection method leveraging language-guided embeddings, and its extension, WinCLIP+, which incorporates few-shot learning [3]. We also include MMR, a reconstruction-based method [12], as a benchmark to compare performance across different data regimes. By focusing on the AeBAD dataset, which introduces significant domain variability, this study aims to bridge the gap between few-shot and fully-supervised anomaly detection methods, providing insights into their robustness and practical applicability.

3 Dataset

The AeBAD dataset is a real-world benchmark specifically designed for anomaly detection in aero-engine blade inspections. It was introduced by Zhang et al. [12], and comprises diverse images of aero-engine blades captured subject to varying domain shifts, such as changes in background, illumination, and viewpoint, pictured in Figure 1. This diversity reflects the operational challenges of anomaly detection in real-world industrial scenarios. MMR, an anomaly detection method was published alongside this dataset serving as a benchmark.

Figure 1: Example non-anomalous images with domain-shift, from left to right: no domain-shift, background, illumination, view.

4 Methodology

In this section, we describe the investigated anomaly detection methods as well as the experimental setup we used for their evaluation.

4.1 Investigated Methods

WinCLIP: Zero-Shot Anomaly Detection. WinCLIP [3] leverages pre-trained vision-language embedding to perform anomaly detection without additional fine-tuning. The model combines a visual encoder, based on CLIP [5], with a semantic representation space to detect anomalies by comparing the embedding with natural language descriptors of normal behaviour. The zero-shot setting enables direct application without the need for any training data, making it well suited for scenarios where annotated anomalies are scarce.

The anomaly detection process involves two key steps. First, visual features are embedded using a backbone model. Second, their cosine similarity is computed against predefined textual prompts representing "normal" behaviour. Samples with low similarity scores are flagged as anomalous. While this approach effectively generalizes across diverse scenarios, its reliance on coarse embeddings may lead to suboptimal performance in fine-grained anomaly localization.

WinCLIP+: Few-Shot Enhancement. To address the aforementioned limitations of zero-shot settings, WinCLIP+ [3] extends WinCLIP by incorporating a few-shot learning mechanism. This enhancement involves fine-tuning the visual-text alignment using a small set of normal samples, commonly 1 to 4 shots. By updating the alignment, WinCLIP+ improves its ability to differentiate anomalies from subtle variations in normal behaviour.

This fine-tuning is achieved through a contrastive learning objective, optimizing the alignment between visual features and corresponding textual embeddings.

In the original paper, this process was found to significantly reduce false positives and enhance anomaly localization.

MMR: Reconstruction Based Benchmark. MMR (Maximum Mean Reconstruction) [12] is included as a benchmark method to evaluate performance against a traditional reconstruction-based approach. MMR uses a generative model to reconstruct the input image. These reconstructions are then used to compute anomaly scores. While MMR is designed for fully-supervised settings, we adapt it to a reduced-data regime to ensure fair comparison with WinCLIP+.

4.2 Experimental Setup

As no official WinCLIP implementation has been published, in our experiments, we used an unofficial implementation developed for benchmarking in a paper by Zhu and Pang [13]. For the investigation of MMR, the official implementation was used. The following paragraphs detail the performed experiments.

WinCLIP: Zero-Shot. : In this experiment, we directly applied WinCLIP without any fine-tuning on the test portion of the dataset.

WinCLIP+: Few-Shot. : To evaluate few-shot capabilities of WinCLIP+, it was fine-tuned with a limited number of normal samples. In particular, we investigated a 1-shot and 4-shot variants.

WinCLIP+: Many-Shot. : For the investigation of the effect of training with a larger amount of data, WinCLIP+ was trained with larger subsets of normal samples. The presented scenarios are 10-shot, 50-shot and 75-shot.

Table 1: Sample-Level AUROC% on the AeBAD dataset. Best values are highlighted.

Source	Method	Same	Background	Illumination	View	Mean
Zhang et al. [12]	PatchCore	75.2 ± 0.3	74.1 ± 0.3	74.6 ± 0.4	60.1 ± 0.4	71.0
	ReverseDistillation	82.4 ± 0.6	84.3 ± 0.9	85.5 ± 0.9	71.9 ± 0.8	81.0
	DRAEM	64.0 ± 0.4	62.1 ± 6.1	61.6 ± 2.7	62.3 ± 0.9	62.5
	NSA	66.5 ± 1.4	48.8 ± 3.5	55.5 ± 3.2	55.9 ± 1.1	56.7
	RIAD	38.6 ± 0.6	41.6 ± 1.3	46.8 ± 0.8	33.0 ± 0.6	40.0
	InTra	39.8 ± 0.8	46.1 ± 0.5	44.7 ± 0.3	46.3 ± 1.5	44.2
Our Findings	MMR	**85.7 ± 0.3**	**84.4 ± 0.7**	**88.7 ± 0.6**	79.5 ± 0.5	**84.6**
	WinCLIP (0-Shot)	80.3 ± 0.2	82.9 ± 0.5	67.0 ± 0.3	82.0 ± 0.3	78.0
	WinCLIP+ (1-Shot)	80.7 ± 0.5	83.1 ± 0.5	67.4 ± 0.6	**82.1 ± 0.4**	78.3
	WinCLIP+ (4-Shot)	80.9 ± 0.2	83.7 ± 0.4	67.7 ± 0.4	81.9 ± 0.3	78.6

Table 2: Sample-Level AUROC% of WinCLIP+ in many-shot settings on the AeBAD dataset.

Method	Same	Background	Illumination	View	Mean
WinCLIP+ (10-Shot)	81.1 ± 0.3	83.7 ± 0.4	67.7 ± 0.4	81.8 ± 0.6	78.6
WinCLIP+ (50-Shot)	81.2 ± 0.1	83.8 ± 0.3	67.4 ± 0.5	81.9 ± 0.6	78.6
WinCLIP+ (75-Shot)	81.2 ± 0.2	84.0 ± 0.2	67.5 ± 0.3	81.9 ± 0.6	78.6

MMR: Full Spectrum. : MMR is the official benchmark for the AeBAD dataset. It is normally trained on the full dataset. To allow for a more detailed comparison, we investigated its performance on a spectrum of possible fractions of training data. This allows us to compare the methods in different use cases, distinguished by the amount of available data. In particular, we trained with 3, 13, 25, 51, 77, 129, 180, 233, 285, 363, 441 and 521 shots. For each of these fractions, we performed 5 experiments with randomly selected seeds.

4.3 Evaluation Metrics

To evaluate sample-level anomaly detection, we use the area under the receiver operating characteristic (AUROC) curve, which is a metric that takes into account both false positives and false negatives and is widely used in the topic of anomaly detection. We report the value as a percentage.

5 Results

In this section, we present the findings of our experiments which are detailed in the previous section.

5.1 WinCLIP & WinCLIP+

Table 1 shows the sample level AUROC% for WinCLIP and WinCLIP+ in their original zero- and few-shot settings.

WinCLIP demonstrates strong generalization capabilities for anomaly classification tasks without requiring any fine-tuning. However, its performance is limited when addressing cases with subtle or localized anomalies, highlighting potential challenges in fine-grained anomaly detection. By incorporating minimal labeled data in the few-shot setting, WinCLIP+ significantly enhances classification performance, effectively addressing domain shifts such as variations in illumination and background.

The incorporation of a small number of normal samples significantly enhances the performance. Between 1 to 4 shots, WinCLIP+ reduces false positives and improves anomaly localization, particularly in scenarios with domain shifts.

Of note is the fact that for the view domain-shift, WinCLIP and WinCLIP+ outperform all other methods.

5.2 WinCLIP+: Many-Shot

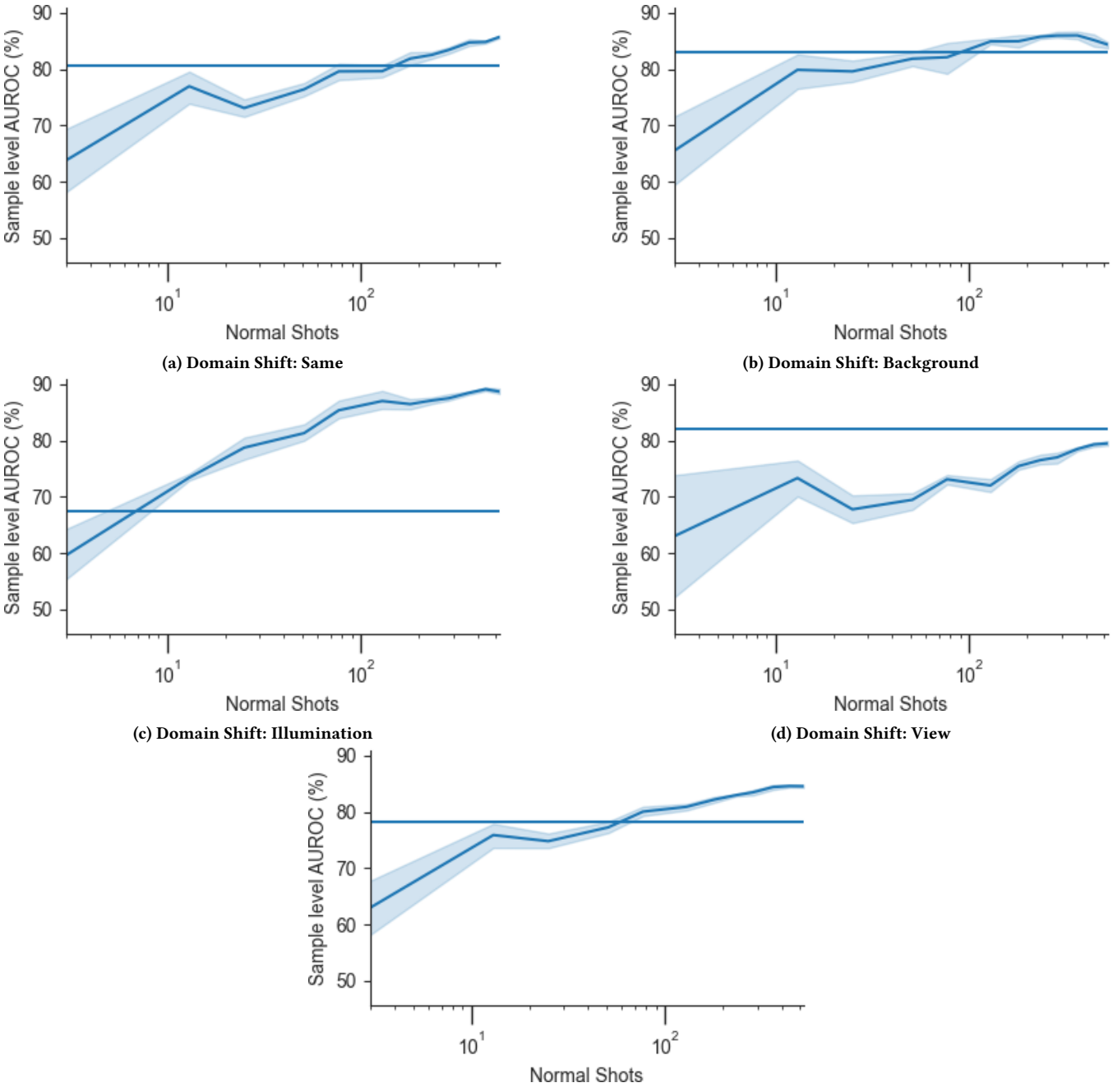
Table 2 presents the performance of WinCLIP+ trained with an increased number of normal-shots. The method's performance improves only slightly, with the mean performance being virtually static. This highlights the main application of WinCLIP+ being low-data regimes.

5.3 MMR: Full Spectrum

From the data presented in Table 1 seems that the best performing method is the original benchmark, MMR. This is true when comparing the methods in their intended data regimes. These regimes have, however, different applications. For a more informed comparison, our investigation, presented in Figure 2, shows a substantial decay of this method's performance with decreasing training dataset size.

For background and no domain-shifts, WinCLIP+ (1-shot) performs comparably to MMR trained on roughly 100 training images.

In the case of the view domain-shift, WinCLIP+ (1-shot) outperforms MMR trained on all fractions of the data. The illumination domain shift, however paints a contrasting picture. In this case, the performance of WinCLIP and all WinCLIP+ variants, also substantiated in Tables 1 and 2, is much worse, with MMR surpassing it with less than 10 training shots. These results demonstrate that WinCLIP+ holds a distinct advantage in low-data regimes.



(a) Domain Shift: Same

(b) Domain Shift: Background

(c) Domain Shift: Illumination

(d) Domain Shift: View

(e) Mean Across Domain Shifts

Figure 2: Sample-Level AUROC% of MMR at different training dataset sizes for each domain-shift, (a) - (d), and the overall mean (e). The light area represents the 95% confidence interval, and the performance of WinCLIP (1-shot) is represented with a horizontal line.

6 Discussion

Another factor differentiating the methods is time. The training and inference of MMR on the full dataset took approximately 160 minutes on one NVIDIA GeForce GTX 1080 Ti with 11GB of memory. Training on 100 shots, just less than one fifth of the data, takes instead around 40 minutes on the same GPU.

In comparison, evaluating WinCLIP (1-shot) required approximately 98 minutes using a single NVIDIA RTX 3060. This evaluation

time reflects both the preprocessing and inference steps needed for the model. Although taking longer than MMR with comparable performance, WinCLIP+'s reliance on minimal labelled data and its ability to generalize well in few-shot settings offsets this computational cost.

The observed trade-off between computation time and data efficiency highlights WinCLIP+'s suitability for real-world scenarios where data collection is expensive and time-consuming. This matters, especially, for application in fields with lower production output, such as aero-blade manufacturing and inspection.

On the other hand, for application in a high-output setting, such as electronics factory producing the same component in huge quantities, model training is required only once. These applications suit more traditional methods, like MMR, thanks to the better speed and performance.

Future work may explore optimization techniques to further reduce this time disparity in time while maintaining the robust performance of few-shot anomaly detection methods.

7 Future Work

Anomaly Localization. In the course of our work, we managed to investigate anomaly classification. Most of the mentioned methods are capable of anomaly localisation, also known as anomaly segmentation. Continuing our experiments in this area of anomaly detection will provide a well rounded method comparison.

Domain-shifts: Our findings show WinCLIP+ struggling with the illumination domain-shift. On the other hand, it excels in the case of the view domain-shift. An in-depth investigation of the mechanics behind this disparity may reveal opportunities for improvement for both WinCLIP+ and MMR.

Training shot ensembles: So far, our work did not entail experiments aimed at improving WinCLIP+'s performance. Among other possible improvement techniques, we are hoping to explore the possibility of the improvement by guiding the focus of the method with purposefully designed training shot set containing multiple domain shifts, not necessarily the ones present in the testing dataset.

Further benchmark few-shot investigation: In our work, we investigated the performance of a non-few-shot method MMR in a few-shot scenario. MMR is not the only well-performing method. Investigating possible performance decay of methods based on different techniques might yield novel insight.

8 Conclusion

In our work, we successfully tested WinCLIP+'s anomaly classification capability in a real-world use case, the AeBAD dataset. We also compared its performance with the current state-of-the-art anomaly detection method, MMR, and showed WinCLIP+'s superiority in a few-shot scenario.

Acknowledgments

We would like to give special thanks to the ETH AI Center and IBM Research Europe for their support during our work.

References

[1] Amy Brand, Liz Allen, Micah Altman, Marjorie Hlava, and Jo Scott. 2015. Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing* 28, 2 (2015), 2.

[2] Geoffrey E Hinton and Ruslan R Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313, 5786 (2006), 504–507.

[3] Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19606–19616.

[4] Ian T Jolliffe. 1986. *Principal component analysis*. Springer.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[6] Karsten Roth et al. 2022. Towards total recall in industrial anomaly detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2022).

[7] Thomas Schlegl et al. 2017. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. *arXiv preprint arXiv:1703.05921* (2017).

[8] Bernhard Schölkopf et al. 1999. Support vector method for novelty detection. *Advances in neural information processing systems* 12 (1999), 582–588.

[9] Jake Snell et al. 2017. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*. 4077–4087.

[10] Flood Sung et al. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1199–1208.

[11] Vid Zavrtanik et al. 2021. DRAEM: A discriminatively trained reconstruction embedding for surface anomaly detection. *Proceedings of the IEEE conference on computer vision and pattern recognition* (2021).

[12] Zilong Zhang, Zhibin Zhao, Xingwu Zhang, Chuang Sun, and Xuefeng Chen. 2023. Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction. *Computers in Industry* 151 (2023), 103990.

[13] Jiawen Zhu and Guansong Pang. 2024. Toward Generalist Anomaly Detection via In-context Residual Learning with Few-shot Sample Prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.